Electronic Journal of Statistics
Vol. 14 (2020) 2892–2921
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1741

Consistent estimation of high-dimensional factor models when the factor number is over-estimated

Matteo Barigozzi

Department of Economics, Università di Bologna
Bologna, Italy
e-mail: matteo.barigozzi@unibo.it

Haeran Cho

School of Mathematics, University of Bristol
Bristol, UK
e-mail: haeran.cho@bristol.ac.uk

Abstract: A high-dimensional r-factor model for an n-dimensional vector time series is characterised by the presence of a large eigengap (increasing with n) between the r-th and the $(r+1)$-th largest eigenvalues of the covariance matrix. Consequently, Principal Component (PC) analysis is the most popular estimation method for factor models and its consistency, when r is correctly estimated, is well-established in the literature. However, popular factor number estimators often suffer from the lack of an obvious eigengap in empirical eigenvalues and tend to over-estimate r due, for example, to the existence of non-pervasive factors affecting only a subset of the series. We show that the errors in the PC estimators resulting from the over-estimation of r are non-negligible, which in turn lead to the violation of the conditions required for factor-based large covariance estimation. To remedy this, we propose new estimators of the factor model based on scaling the entries of the sample eigenvectors. We show both theoretically and numerically that the proposed estimators successfully control for the over-estimation error, and investigate their performance when applied to risk minimisation of a portfolio of financial time series.

MSC2020 subject classifications: Primary 62M10; secondary 62H12.
Keywords and phrases: Factor models, principal component analysis, sample eigenvectors, factor number.

Received November 2019.

1. Introduction

Factor modelling is a popular approach to dimension reduction in high-dimensional time series analysis. It has been successfully applied to large panels of time series for forecasting macroeconomic variables (Stock and Watson, 2002a), building low-dimensional indicators of the whole economic activity (Stock and Watson, 2002b) and analysing dynamic brain connectivity using high-dimensional fMRI data (Ting et al., 2017), to name a few.

In this paper, we consider one of the most general factor models in the literature, the approximate dynamic factor model, which permits serial dependence in the factors and both serial and cross-sectional dependence among the idiosyncratic components. More specifically, given an n-dimensional vector time series $\{\mathbf{x}_t = (x_{1t}, \ldots, x_{nt})^\top, 1 \leq t \leq T\}$, we investigate the problem estimating the factor model

$$x_{it} = \boldsymbol{\lambda}_i^\top \mathbf{f}_t + \varepsilon_{it}, \tag{1}$$

where $\boldsymbol{\lambda}_i$ and $\mathbf{f}_t$ are r-dimensional vectors of loadings and factors, respectively. We refer to $\chi_{it} = \boldsymbol{\lambda}_i^\top \mathbf{f}_t$ as the common component and ε_{it} as the idiosyncratic component, and assume the number of factors, r, to be fixed independent of n and T.

The main assumption that guarantees the asymptotic identification under (1) is the existence of a *large* (increasing with n) eigengap between the r leading eigenvalues of the covariance matrix of $\mathbf{x}_t$ and the remaining ones. Intuitively, since the eigengap is assumed to increase with n, the more series are pooled together, the more the contribution of the factors to the total co-variation in the data is likely to emerge over the idiosyncratic components ('blessing of dimensionality'). As a consequence, a natural way of estimating (1) is via Principal Component (PC) analysis, through which the common components are estimated as the projection of the data onto the space spanned by the leading eigenvectors of the sample covariance matrix, i.e., given some estimator $\widehat{r}$ of the factor number r, the PC estimator of the common component is defined as

$$\widehat{\chi}_{it}^{\mathrm{pc}} = \sum_{j=1}^{\widehat{r}} \widehat{w}_{x,ij} \widehat{\mathbf{w}}_{x,j}^\top \mathbf{x}_t, \tag{2}$$

where $\widehat{\mathbf{w}}_{x,j} = (\widehat{w}_{x,1j}, \ldots, \widehat{w}_{x,nj})^\top$ is the normalised eigenvector corresponding to the j-th largest eigenvalue of the sample covariance matrix of $\mathbf{x}_t$. The PC estimator (2) allows for consistent estimation of the common component of model (1), provided that both $n, T \to \infty$ (see Bai, 2003, and Fan, Liao and Mincheva, 2013).

However, the theoretical properties of PC estimators have always been investigated conditional on $\widehat{r}$ being a consistent estimator of r, and the problem of determining r has typically been treated separately. Many methods exist for estimating the factor number: Bai and Ng (2002), Alessi, Barigozzi and Capasso (2010), see Onatski (2010), Ahn and Horenstein (2013), Yu, He and Zhang (2018), Trapani (2018), and Bai and Ng (2019), to name a few, all of which exploit the postulated existence of the eigengap. On the other hand, it is often difficult to identify the large gap from empirical eigenvalues. In particular, it is known that the presence of moderate cross-sectional correlations in the idiosyncratic components shrinks the empirical eigengap by introducing some so-called 'weak' factor (Onatski, 2012), and we empirically demonstrate that commonly adopted factor number estimators often over-estimate r in such situations. Moreover, as noted in Barigozzi, Cho and Fryzlewicz (2018), instabilities in the factor structure tend to spuriously enlarge the factor space and

introduce further difficulties to determining the number of factors. Finally, as shown later in the paper, different estimators frequently return discordant results, thus making it ambiguous for the user to choose a single value to rely on.

1.1. Our contributions

The question is, what do we do if we have a range of possible candidate estimators of r, or if we believe that none of the estimators is reliable? One may use the largest number of factors returned by available methods, or set it to be even larger, with the expectation of avoiding the hazard of under-estimating the factor-driven variation, which is a problem without any clear solution.

In this paper, we first show that over-estimation of r can incur non-negligible estimation error when considering the worst case scenarios for individual common components (see Proposition 2). To the best of our knowledge, this problem has not been investigated in the literature before. Identifying the theoretical difficulties arising under the time series factor model, we propose a novel blockwise estimation technique that enables rigorous treatment of the PC-based estimators which is another contribution made in this paper.

In order to mitigate the lack of a reliable estimator of r, we propose a modified PC estimator which performs as well as the 'oracle' estimator constructed *with the knowledge of true* r and, consequently, makes our estimation procedure free from the difficult task of estimating r accurately.

More specifically, the factor model (1) is usually characterised by the following eigengap conditions (see e.g. Fan, Liao and Mincheva, 2013):

(C1) there exist some fixed $\underline{c}_j, \bar{c}_j$ such that for $1 \le j \le r$,

$$0 < \underline{c}_j < \liminf_{n\to\infty} \frac{\mu_{\chi,j}}{n} \le \limsup_{n\to\infty} \frac{\mu_{\chi,j}}{n} < \bar{c}_j < \infty$$

and $\bar{c}_{j+1} < \underline{c}_j$ for $j \le r-1$,

(C2) $\mu_{\varepsilon,1} < C_\varepsilon < \infty$ for any n,

where $\mu_{\chi,j}$ and $\mu_{\varepsilon,j}$ denote the j-th largest eigenvalues of the covariance matrices of the common and idiosyncratic components, respectively. The linear divergence of eigenvalues in (C1) is a prevailing and natural assumption in the factor model literature, implying that all series in the panel are equally important for the recovery of the factors. From (C1), it follows that $\mathbf{w}_{\chi,j} = (w_{\chi,1j}, \ldots, w_{\chi,nj})^\top$, the normalised eigenvector of the covariance matrix of $\boldsymbol{\chi}_t$ corresponding to $\mu_{\chi,j}$, has its coordinates asymptotically bounded as $\max_{1\le i\le n} |w_{\chi,ij}| = O(n^{-1/2})$ for all $j \le r$ (see (6) below). Thanks to the eigengap and the Davis-Kahan theorem (Yu, Wang and Samworth, 2015), the coordinates of the r leading eigenvectors of the sample covariance matrix of the data, $\widehat{\mathbf{w}}_{x,j}$, $j \le r$, are also bounded asymptotically as $\max_{1\le j\le r} \max_{1\le i\le n} |\widehat{w}_{x,ij}| = O_p(n^{-1/2})$. On the other hand, precisely due to the lack of this eigengap, meaningful control of the behaviour of $\widehat{w}_{x,ij}$, $j \ge r+1$ is not obvious under the dynamic factor model in (1).

Motivated by these observations, we propose to modify $\widehat{\mathbf{w}}_{x,j}$ via *scaling* as

$$\widehat{\mathbf{w}}^{\mathrm{sc}}_{x,j} = \nu_j^{-1}\,\widehat{\mathbf{w}}_{x,j} \quad \text{with} \quad \nu_j = \max\{1, \delta_n^{-1} \max_{1\le i\le n} |\widehat{w}_{x,ij}|\}, \tag{3}$$

which ensures that the entries of the modified eigenvectors are bounded by some δ_n of order $n^{-1/2}$. By substituting $\widehat{\mathbf{w}}^{\mathrm{sc}}_{x,j}$ in place of $\widehat{\mathbf{w}}_{x,j}$ in (2), we obtain a novel *scaled* PC estimator of the common component. While conceptually and computationally simple, the scaled PC estimator attains the same asymptotic error bound as the oracle PC estimator obtained with the true r, successfully curtailing the error attributed to spurious factors without requiring the accurate estimation of the factor number beyond that $\widehat{r} \ge r+1$. We also propose a well-motivated choice of the tuning parameter δ_n.

The good performance of the scaled PC estimator when r is over-estimated, in contrast to that of the PC estimator, is demonstrated on simulated datasets. In addition, we investigate the impact of the non-negligible errors in the PC estimator (or lack thereof in the modified PC estimator) on large covariance matrix estimation through an application to risk minimisation of a portfolio of financial time series.

1.2. Relationship to the existing literature

Recently, Bai and Ng (2019) adopted the eigenvalue shrinkage for minimum-rank factor analysis under time series factor models. Our approach is distinguished from theirs in that we aim at avoiding the reliance on the accurate estimation of the factor number itself in establishing the theoretical consistency of the estimator of common components. Sharing the aim closer to ours, Fan and Liao (2019) propose a diversified factor estimator obtained as cross-sectional averages of the data with respect to pre-determined weights and show their robustness to over-estimating the number of factors.

We mention two other approaches to time series factor analysis for which our work can be relevant. First, assuming that all serial dependence in the data is captured by the factors, Lam, Yao and Bathia (2011) and Lam and Yao (2012) proposed an alternative approach to factor model analysis. Since their method is also based on eigenanalysis of a suitable covariance matrix, our methodology can be readily adapted to this case as well. Second, Forni et al. (2000) considered a richer factor structure where factors are allowed to have lagged effects on the data. Estimation of such model is in general based on spectral PC analysis, but other approaches exist that require standard PC analysis at the initial or final step (e.g., Forni et al., 2005, Bai and Ng, 2007, Forni et al., 2009, and Doz, Giannone and Reichlin, 2011), and our proposed modifications can be easily adopted for this purpose.

Finally, we note that there are some links between the model and the estimators proposed here and the vast literature on statistical models and methods based on random matrix theory, see El Karoui (2008), Cai, Ma and Wu (2013), Donoho, Gavish and Johnstone (2018) and Donoho and Ghorbani (2018), and also Paul and Aue (2014) for an overview.

Structure of the paper

The rest of the paper is organised as follows. We introduce the approximate factor model in Section 2, where we also discuss its estimation via PC, and we investigate the behaviour of factor number estimators as well as the impact of over-estimating the factor number on the PC estimator. In Section 3, we motivate and introduce the modified PC estimator based on scaling, and study its theoretical properties. Comparative simulation study of PC-based estimators is conducted in Section 4, and we apply the proposed estimators to financial data analysis in Section 5. All the proofs of the main theoretical results are in Appendix A. An extended version of this manuscript containing additional theoretical and simulation results is available as Barigozzi and Cho (2020).

Notation

For a given $m \times n$ matrix $\mathbf{B}$ with b_{ij} denoting its (i,j) element, its spectral norm is defined as $\|\mathbf{B}\| = \sqrt{\mu_1(\mathbf{B}\mathbf{B}^\top)}$, where $\mu_k(\mathbf{C})$ denotes the k-th largest eigenvalue of $\mathbf{C}$, its Frobenius norm as $\|\mathbf{B}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n b_{ij}^2}$, and also $\|\mathbf{B}\|_{\max} = \max_{1\le i\le m} \max_{1\le j\le n} |b_{ij}|$. The sub-exponential norm of a random variable X is defined as $\|X\|_{\psi_1} = \inf_k\{k : \mathsf{E}[\exp(|X|/k)] \le 2\}$. For a given set Π, we denote its cardinality by $|\Pi|$. For any vector $\mathbf{a} = (a_1, \dots, a_m) \in \mathbb{R}^m$, we denote $\|\mathbf{a}\|_0 = |\{1 \le i \le m : a_i \ne 0\}|$ and $\|\mathbf{a}\|_\infty = \max_{1\le i\le m} |a_i|$. Also, we use the notations $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$. The notation $a \asymp b$ indicates that a is of the order of b, and $a \gg b$ indicates that $a^{-1}b \to 0$. We denote an $m \times m$-identity matrix by $\mathbf{I}_m$.

2. The approximate dynamic factor model

2.1. Model and assumptions

Recall the factor model in (1), where an n-dimensional vector time series $\mathbf{x}_t = (x_{1t}, \dots, x_{nt})^\top$ is divided into the common component $\boldsymbol{\chi}_t = (\chi_{1t}, \dots, \chi_{nt})^\top = \mathbf{\Lambda}\mathbf{f}_t$ driven by the vector of r latent factors $\mathbf{f}_t = (f_{1t}, \dots, f_{rt})^\top$, with $\mathbf{\Lambda} = [\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_n]^\top$ as the $n \times r$ matrix of loadings, and the idiosyncratic component $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \dots, \varepsilon_{nt})^\top$. Without loss of generality, we assume $\mathsf{E}(f_{jt}) = \mathsf{E}(\varepsilon_{it}) = 0$ for all i, j, t.

We now list and motivate the assumptions imposed on the approximate dynamic factor model (1) (see e.g., Fan, Liao and Mincheva (2013) and Barigozzi, Cho and Fryzlewicz (2018) for similar conditions).

Assumption 1 (Identification)**.**

(i) $\mathsf{E}(\mathbf{f}_t\mathbf{f}_t^\top) = \mathbf{I}_r$ for all $t \ge 1$.

(ii) There exists a positive definite $r \times r$ matrix $\mathbf{H}$ with distinct eigenvalues and such that $n^{-1}\mathbf{\Lambda}^\top\mathbf{\Lambda} \to \mathbf{H}$ as $n \to \infty$.

(iii) There exists $\bar{\lambda} \in (0, \infty)$ such that $\|\mathbf{\Lambda}\|_{\max} < \bar{\lambda}$.

(iv) There exists $C_\varepsilon \in (0, \infty)$ such that, for any $t \geq 1$,

$$\sum_{i=1}^{n} \sum_{i'=1}^{n} a_i a_{i'} \mathsf{E}(\varepsilon_{it} \varepsilon_{i't}) < C_\varepsilon$$

for any sequence of coefficients $\{a_i\}_{i=1}^n$ satisfying $\sum_{i=1}^n a_i^2 = 1$.

(v) $\mathsf{E}(f_{jt} \varepsilon_{it'}) = 0$ for all $i \leq n$, $j \leq r$ and $t, t' \leq T$.

We adopt the normalisation given in Assumption 1 (i)–(ii) for the purpose of identification; in general, factors and loadings are recoverable up to a linear invertible transformation only. Assumption 1 (iii) is a commonly found assumption in the factor model literature (see Assumption B in Bai (2003)) which, together with Assumption 1 (ii), requires that factors influence all cross-sections to a similar degree. Assumption 1 (iv) allows for mild cross-sectional dependence across idiosyncratic components. In other words, we are considering an *approximate* factor structure, as opposed to the classical *exact* factor model where $\boldsymbol{\varepsilon}_t$ is assumed to be uncorrelated cross-sectionally. It is possible to relax Assumption 1 (v) and allow for weak dependence between the factors and the idiosyncratic components (c.f. Assumption D of Bai and Ng, 2002).

In order to motivate the assumptions further, we adopt the notations

$$\boldsymbol{\Gamma}_\chi = \boldsymbol{\Lambda} \left(\frac{1}{T} \sum_{t=1}^{T} \mathsf{E}(\mathbf{f}_t \mathbf{f}_t^\top) \right) \boldsymbol{\Lambda}^\top = \boldsymbol{\Lambda} \boldsymbol{\Lambda}^\top, \quad \boldsymbol{\Gamma}_\varepsilon = \frac{1}{T} \sum_{t=1}^{T} \mathsf{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^\top), \quad \text{and}$$
$$\boldsymbol{\Gamma}_x = \boldsymbol{\Gamma}_\chi + \boldsymbol{\Gamma}_\varepsilon.$$

If $\mathbf{f}_t$ and $\boldsymbol{\varepsilon}_t$ are covariance stationary, then these matrices are the corresponding population covariance matrices. Also, we denote the eigenvalues (in non-increasing order) of $\boldsymbol{\Gamma}_\chi$, $\boldsymbol{\Gamma}_\varepsilon$ and $\boldsymbol{\Gamma}_x$ by $\mu_{\chi,j}$, $\mu_{\varepsilon,j}$ and $\mu_{x,j}$, respectively. Then, Assumption 1 leads to (C1)–(C2) in Section 1.1, i.e., $\mu_{\chi,j}$, $j \leq r$ diverge linearly in n as $n \to \infty$, whereas $\mu_{\varepsilon,1}$ is bounded for any n. This condition coincides with Definition 2 in Chamberlain and Rothschild (1983) and Assumption 2 in Fan, Liao and Mincheva (2013), and it is also in the same spirit as Assumption C.4 in Bai (2003) where cross-sectional dependence of idiosyncratic components is assumed to be weak.

Moreover, (C1)–(C2) imply that, due to Weyl's inequality, the eigenvalues of $\boldsymbol{\Gamma}_x$, $\mu_{x,j}$, satisfy the following eigengap conditions:

(C3) The r largest eigenvalues, $\mu_{x,1}, \ldots, \mu_{x,r}$, diverge linearly in n as $n \to \infty$;
(C4) the $(r+1)$-th largest eigenvalue, $\mu_{x,r+1}$, stays bounded for any n.

From (C1)–(C4) above, it is clear that for consistent estimation of the common components, approximate factor models need to be considered in the asymptotic limit where $n \to \infty$, i.e., these models enjoy what is sometimes referred to as the blessing of dimensionality. In particular, we require:

Assumption 2. $n \to \infty$ as $T \to \infty$, with $n = O(T^\kappa)$ for some $\kappa \in (0, \infty)$.

Under Assumption 2, we operate in a high-dimensional setting that permits $n \gg T$, unlike in the random matrix theory literature where it is typically assumed that $n/T \to y \in (0, \infty)$ (Johnstone, 2001). Furthermore we assume:

Assumption 3 (Tail behaviour)**.**
(i) $\max_{1\le j\le r} \max_{1\le t\le T} \|f_{jt}\|_{\psi_1} < B_f$ for some $B_f \in (0, \infty)$.
(ii) $\max_{1\le t\le T} \|\boldsymbol{\varepsilon}_t\|_{\psi_1} < B_\varepsilon$ for some $B_\varepsilon \in (0, \infty)$, where $\|\boldsymbol{\varepsilon}_t\|_{\psi_1} = \sup_{\mathbf{v}\in\mathbb{R}^n:\,\|\mathbf{v}\|=1} \|\mathbf{v}^\top \boldsymbol{\varepsilon}_t\|_{\psi_1}$.

Assumption 4 (Strong mixing)**.** Denoting the σ-algebra generated by $\{(\mathbf{f}_t, \boldsymbol{\varepsilon}_t), s \le t \le e\}$ by $\mathcal{F}_s^e$, let $\alpha(k) = \max_{1\le t\le T} \sup_{A\in\mathcal{F}_{-\infty}^t, B\in\mathcal{F}_{t+k}^\infty} |\mathsf{P}(A)\mathsf{P}(B) - \mathsf{P}(A \cap B)|$. Then, there exist some fixed $c_\alpha, \beta \in (0, \infty)$, such that $\alpha(k) \le \exp(-c_\alpha k^\beta)$ for all $k, T \in \mathbb{Z}^+$.

The sub-exponential tail conditions in Assumption 3, along with Assumption 4, allow us to control the deviation of sample covariance estimates from their population counterparts via Bernstein-type inequality (see Theorem 1 of Merlevède, Peligrad and Rio, 2011) under the approximate dynamic factor model. We stress that either strict or weak stationarity of f_{jt} and ε_{it} is not required in performing the PC-based estimation, provided that the loadings are time-invariant.

2.2. Estimation via principal component analysis

The most common way to estimate the approximate factor model (1) is by means of PC analysis, and the asymptotic properties of the PC estimator have been well-established: in particular, we refer to Fan, Liao and Mincheva (2013) where a set-up similar to ours is considered.

Recall that the PC estimator of the common component: $\widehat{\chi}_{it}^{\mathrm{pc}} = \sum_{j=1}^{\widehat{r}} \widehat{w}_{x,ij} \widehat{\mathbf{w}}_{x,j}^\top \mathbf{x}_t$, where $\widehat{\mathbf{w}}_{x,j}$ denote the j-th leading normalised eigenvector of the sample covariance $\widehat{\boldsymbol{\Gamma}}_x = T^{-1} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t^\top$, and $\widehat{r}$ is an estimator of the number of factors r. Theorem 1 of Barigozzi, Cho and Fryzlewicz (2018), which is a refinement of Corollary 1 of Fan, Liao and Mincheva (2013), establishes a uniform bound on the estimation error over $1 \le i \le n$ and $1 \le t \le T$ of the PC estimator when r is *known*, i.e., $\widehat{r} = r$, under Gaussianity of the idiosyncratic component. Here, we generalise the theorem to the case of sub-exponential distributions as specified in Assumption 3. Its proof can be found in Section B.2 of Barigozzi and Cho (2020).

Proposition 1. *Under Assumptions 1–4, the PC estimator $\widehat{\chi}_{it}^{\mathrm{pc}}$ with $\widehat{r} = r$ satisfies*

$$\max_{1\le i\le n} \max_{1\le t\le T} |\widehat{\chi}_{it}^{\mathrm{pc}} - \chi_{it}| = O_p\left\{ \left(\sqrt{\frac{\log(n)}{T}} \vee \frac{1}{\sqrt{n}} \right) \log(T) \right\}.$$

Two key results are required for proving Proposition 1. First, we make use of the eigengap between $\mu_{x,r}$ and $\mu_{x,r+1}$ increasing linearly in n (see (C3)–(C4)), which ensures that the eigenspace of $\boldsymbol{\Gamma}_\chi$ is consistently estimated by the

r leading eigenvectors of $\widehat{\mathbf{\Gamma}}_x$. More specifically, there exists a diagonal $r \times r$ matrix $\mathbf{S}$ with entries ± 1, such that

$$\|\widehat{\mathbf{W}}_x - \mathbf{W}_\chi \mathbf{S}\| \leq \frac{2^{3/2}\sqrt{r}\|\widehat{\mathbf{\Gamma}}_x - \mathbf{\Gamma}_\chi\|}{\mu_{\chi,r}} = O_p\left(\sqrt{\frac{\log(n)}{T}} \vee \frac{1}{n}\right), \tag{4}$$

where $\widehat{\mathbf{W}}_x = [\widehat{\mathbf{w}}_{x,j},\, j \leq r]$ and $\mathbf{W}_\chi = [\mathbf{w}_{\chi,j},\, j \leq r]$. The result in (4) follows from the modified Davis-Kahan theorem of Yu, Wang and Samworth (2015), the lower bound of $\underline{c}_r n$ on $\mu_{\chi,r}$ from (C1), and the closeness between $\widehat{\mathbf{\Gamma}}_x$ and $\mathbf{\Gamma}_x$ under Assumptions 2–4 (see Lemma 3 (i) in Section A.1). We can further show that

$$\sqrt{n}\left\|\boldsymbol{\varphi}_i^\top(\widehat{\mathbf{W}}_x - \mathbf{W}_\chi \mathbf{S})\right\| = O_p\left(\sqrt{\frac{\log(n)}{T}} \vee \frac{1}{\sqrt{n}}\right), \tag{5}$$

where $\boldsymbol{\varphi}_i$ an n-vector of zeros except for its i-th element being one, see Lemma 3 (iii).

Secondly, denoting the eigendecomposition of the covariance matrix of the common components by $\mathbf{\Gamma}_\chi = \mathbf{W}_\chi \mathbf{M}_\chi \mathbf{W}_\chi^\top$ with $\mathbf{M}_\chi = \mathsf{diag}(\mu_{\chi,1}, \ldots, \mu_{\chi,r})$, (C1) leads to

$$\max_{1\leq i\leq n}\sqrt{\sum_{j=1}^{r} w_{\chi,ij}^2} = \max_{1\leq i\leq n}\|\boldsymbol{\varphi}_i^\top \mathbf{W}_\chi\| \leq \max_{1\leq i\leq n}\|\boldsymbol{\varphi}_i^\top \mathbf{\Gamma}_\chi\|\,\|\mathbf{W}_\chi\|\,\|\mathbf{M}_\chi^{-1}\| = O\left(\frac{1}{\sqrt{n}}\right), \tag{6}$$

i.e., asymptotically, each element of $\mathbf{W}_\chi$ is $O(n^{-1/2})$. This, combined with (5), leads to

$$\max_{1\leq i\leq n}\max_{1\leq j\leq r}|\widehat{w}_{x,ij}| = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{7}$$

The bound in (7) serves as the main motivation behind introducing the modified PC estimators in Section 3.

Remark 1 (Optimality of PC)**.** The PC estimator is appealing for the following reasons. First, under the assumption of spherical idiosyncratic components, $\boldsymbol{\varepsilon}_t \sim_{\text{iid}} \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ for some $\sigma^2 > 0$, the PC estimator of the loadings is asymptotically equivalent to their Maximum Likelihood estimator (Tipping and Bishop, 1999). Second, the sample principal subspace estimator is minimax rate optimal, see Theorem 5 of Cai, Ma and Wu (2013) which shows that $\mathsf{E}\|\widehat{\mathbf{W}}_x\widehat{\mathbf{W}}_x^\top - \mathbf{W}_x\mathbf{W}_x^\top\|_F^2 \asymp rn/(\mu_{\chi,r}T)$. This, combined with (C1), is comparable to the convergence rate reported in (4), although the latter is obtained under the more general approximate dynamic factor model. Third, when allowing for non-spherical and possibly correlated idiosyncratic components, the PC estimator by definition delivers the linear combination of the data with largest variance in the sense that, for the j-th PC, $\widehat{\mathsf{Var}}(\widehat{\mathbf{w}}_{x,j}^\top \mathbf{x}_t) \geq \widehat{\mathsf{Var}}(\boldsymbol{\omega}^\top \mathbf{x}_t)$ for any $\boldsymbol{\omega}$ satisfying $\|\boldsymbol{\omega}\| = 1$ and $\boldsymbol{\omega}^\top \widehat{\mathbf{w}}_{x,j'} = 0$ for any $j' \leq j-1$, where $\widehat{\mathsf{Var}}(\cdot)$ denotes the sample variance operator.

2.3. (Over-)estimation of the number of factors

In practice, the true number of factors r is unknown and its estimation has been one of the most researched problems in the factor model literature (see the references in the Introduction). Based on the conditions (C3)–(C4), a prevailing approach is to identify a 'large' gap between the successive estimated eigenvalues $\widehat{\mu}_{x,j}$, $1 \le j \le r_{\max}$ of the sample covariance matrix $\widehat{\mathbf{\Gamma}}_x$, where $r_{\max}$ denotes the maximum allowable number of factors often required as an input parameter to the estimation procedure. Here we focus on two of the most popular methods.

The information criterion-based method proposed by Bai and Ng (2002) estimates r as

$$\widehat{r} = \underset{1\le q\le r_{\max}}{\arg\min}\ \mathrm{IC}(q), \ \text{where } \mathrm{IC}(q) = \log\left(\frac{1}{n}\sum_{j=q+1}^{n}\widehat{\mu}_{x,j}\right) + q\cdot g(n,T), \tag{8}$$

with a penalty function $g(n,T)$ satisfying $g(n,T)\to 0$ and $\{(n\wedge T)\cdot g(n,T)\}\to\infty$ as $n,T\to\infty$. The eigenvalue ratio-based estimator by Ahn and Horenstein (2013), returns

$$\widehat{r} = \underset{1\le q\le r_{\max}}{\arg\max}\ \mathrm{GR}(q), \ \text{where } \mathrm{GR}(q) = \frac{\log(1+\widehat{\mu}^*_{x,q})}{\log(1+\widehat{\mu}^*_{x,q+1})} \ \text{with } \widehat{\mu}^*_{x,q} = \frac{\widehat{\mu}_{x,q}}{\sum_{j=q+1}^{n}\widehat{\mu}_{x,j}}. \tag{9}$$

Implicitly, the information criterion in (8) performs thresholding on the scaled sample eigenvalues $\widehat{\mu}^*_{x,q}$ with respect to $g(n,T)$, and selects an index q among those that correspond to $\widehat{\mu}^*_{x,q}$ surviving the thresholding. On the other hand, the eigenvalue ratio approach in (9) considers the ratio of the successive scaled eigenvalues without taking into account the size of the eigenvalues. This difference frequently leads to distinct estimators from the different approaches, not to mention that, as shown in Alessi, Barigozzi and Capasso (2010), the various choices of $g(n,T)$ often result in different factor number estimators. Another parameter whose choice may affect the estimation result for both of the estimators (8)–(9) is $r_{\max}$. Moreover, while (C3)–(C4) are asymptotic conditions, the lack of an obvious eigengap in the empirical eigenvalues poses a challenge in the estimation of r. Consequently, the estimated number of factors is highly variable as the following quantities vary: the dimensions n and T, the degree of cross-sectional correlations in the idiosyncratic components, and the signal-to-noise ratio represented by the ratio between $\mathsf{Var}(\chi_{it})$ and $\mathsf{Var}(\varepsilon_{it})$, see e.g., the numerical studies in Ahn and Horenstein (2013) and Trapani (2018).

For an illustration, we conduct a comparative simulation study by applying the two estimators (8) (with $g(n,T) = (n+T)\log(n\wedge T)/(nT)$, i.e., IC_2 of Bai and Ng (2002)) and (9) with a generous but reasonable choice $r_{\max} = [\sqrt{n\wedge T}]$, to datasets simulated under Model 1 as described in Section 4. The results are reported in Figure 1. It is apparent that the estimator (8) fails to return the true number of factors $r = 5$ in the presence of moderate degree of cross-sectional correlations in $\boldsymbol{\varepsilon}_t$, especially when n is small. While (9) performs considerably better

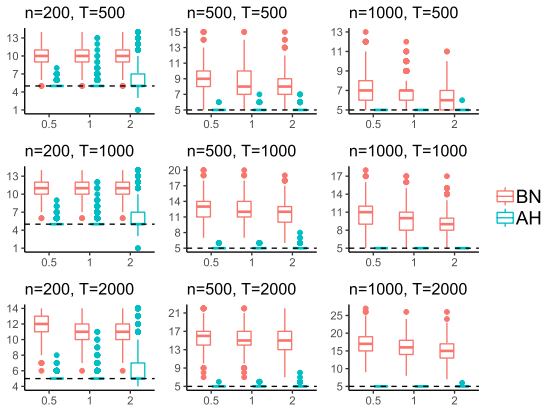

FIG 1. *Box plots of* $\widehat{r}$ *returned by* (8) *(BN) and* (9) *(AH) over* 1000 *realisations generated under Model 1 with* $T \in \{500, 1000, 2000\}$ *(top to bottom),* $n \in \{200, 500, 1000\}$ *(left to right) and* $\phi \in \{0.5, 1, 2\}$ *(left to right within each plot, controls the noise-to-signal ratio); horizontal broken lines indicate the true factor number* $r = 5$.

for this particular data generating process, we provide in Section C.1 of Barigozzi and Cho (2020) the scenarios where this method also fails to return the correct number of factors. We note that the performance of the estimators does not improve with increasing sample size T. In almost all cases considered, the factor number is over-estimated, i.e., $\widehat{r} \geq r$, with (9) occasionally delivering $\widehat{r} < r$.

Obviously, when $\widehat{r} < r$, the PC estimator (2) or indeed, any estimator of the common component does not capture the contribution from one or more factors, which inevitably incurs a non-negligible error and no remedy to this problem exists. To circumvent this problem, the user may be tempted to increase $\widehat{r}$ based on the reasoning that the contribution of spurious factors beyond the r-th one is negligible and thus such a strategy would be risk averse. However, this reasoning is incorrect as we show in the formal theoretical treatment of the impact of over-estimation of r on factor analysis in the next section.

2.4. PC estimator when r is over-estimated

While Onatski (2015) shows in his Proposition 1 that the errors due to the over-estimation of r is negligible once aggregated over cross-sections and time, a formal analysis of the impact of the over-estimated factor number on the PC estimators of *individual* common components has not yet been conducted to the best of our knowledge.

Recalling the PC estimator (2), we have the following decomposition of the estimation error when $\widehat{r} > r$,

$$\widehat{\chi}_{it}^{\mathrm{pc}} - \chi_{it} = \left(\sum_{j=1}^{r} \widehat{w}_{x,ij} \widehat{\mathbf{w}}_{x,j}^{\top} \mathbf{x}_t - \chi_{it} \right) + \sum_{j=r+1}^{\widehat{r}} \widehat{w}_{x,ij} \widehat{\mathbf{w}}_{x,j}^{\top} \mathbf{x}_t. \tag{10}$$

The rate of convergence for the error in the oracle PC estimator (first term in the RHS of (10)) is given in Proposition 1. Our interest lies in the theoretical treatment of the second term representing the over-estimation error. This faces two main challenges.

(a) The large eigengap between $\mu_{\chi,r}$ and $\mu_{\chi,r+1} = 0$ (see (C1)) and Davis-Kahan theorem play a key role in controlling the distance between the empirical principal subspace spanned by the r leading eigenvectors of $\widehat{\mathbf{\Gamma}}_x$ and those of $\mathbf{\Gamma}_\chi$, as reported in (4). On the other hand, due to the lack of eigengap between the successive $\mu_{x,j}$, $j \geq r+1$ (see (C4)) or any other structural assumptions, the behaviour of $\widehat{\mathbf{w}}_{x,j}$ for $j \geq r+1$ cannot be controlled in a meaningful way.

(b) The eigenvectors $\widehat{\mathbf{w}}_{x,j}$, $1 \leq j \leq \widehat{r}$, are obtained from the full sample covariance matrix and thus are dependent on $\mathbf{x}_t$, $1 \leq t \leq T$, which, together with the issue noted in (a), makes it difficult to analyse the stochastic properties of $\widehat{\mathbf{w}}_{x,j}^\top \mathbf{x}_t$ for $j \geq r+1$.

With these difficulties, we derive the following uniform but uninformative upper bound on the over-estimation error:

$$\max_{1\leq i\leq n}\max_{1\leq t\leq T}\left|\sum_{j=r+1}^{\widehat{r}}\widehat{w}_{x,ij}\widehat{\mathbf{w}}_{x,j}^\top\mathbf{x}_t\right| \leq \sum_{j=r+1}^{\widehat{r}}\max_{1\leq i\leq n}|\widehat{w}_{x,ij}|\|\widehat{\mathbf{w}}_{x,j}\|\cdot\max_{1\leq t\leq T}\|\mathbf{x}_t\| = O_p(\sqrt{n}\log(T)). \tag{11}$$

In the next section, we propose modifications of the PC estimator which directly address the issue raised in (a) but first, we introduce a novel 'blockwise' estimation technique which, under the time series factor model (1), allows for bypassing the issue raised in (b) and hence enables a rigorous theoretical analysis of the PC estimator when $\widehat{r} \geq r+1$. For this, we split the data into blocks of size b_T, say $\{\mathbf{x}_t, t \in I_\ell\}$ for $I_\ell := \{(\ell-1)b_T+1, \ldots, \min(\ell b_T, T)\}$, $\ell = 1, \ldots, L_T := \lceil T/b_T \rceil$. Also, denote by $\bar{I}_\ell := \{1, \ldots, T\} \setminus \bigcup_{m\in\{\ell,\ell\pm1\}} I_\ell$, i.e., the set of indices that do not belong to I_ℓ or its adjacent blocks, and by $\widehat{\mathbf{w}}_{x,j}^{(\ell)}$ the j-th leading eigenvector of $\widehat{\mathbf{\Gamma}}_x^{(\ell)} = |\bar{I}_\ell|^{-1}\sum_{t\in\bar{I}_\ell}\mathbf{x}_t\mathbf{x}_t^\top$, i.e., the sample covariance matrix constructed by *omitting* the ℓ-th and its adjacent blocks. Then, we obtain the blockwise PC estimator of χ_{it} as

$$\widehat{\chi}_{it}^{\text{bpc}} = \sum_{j=1}^{\widehat{r}}\widehat{w}_{x,ij}^{(\ell)}(\widehat{\mathbf{w}}_{x,j}^{(\ell)})^\top\mathbf{x}_t \quad \text{for} \quad t \in I_\ell, \quad 1 \leq \ell \leq L_T. \tag{12}$$

In other words, the common components are estimated in a blockwise manner as projections of $\mathbf{x}_t$ onto the principal subspace of the subsample obtained from omitting the current block as well as its immediate neighbours. We select the block size b_T to balance between avoiding the asymptotic loss in efficiency by having $|\bar{I}_\ell| \geq T - 3b_T$ as large as possible, and ensuring that the dependence between $\widehat{\mathbf{w}}_{x,j}^{(\ell)}$ and $\mathbf{x}_t$, $t \in I_\ell$ is sufficiently weak under the strong mixing condi-

tion in Assumption 4 (ii), hence permitting the rigorous theoretical treatment of $(\widehat{\mathbf{w}}_{x,j}^{(\ell)})^\top \mathbf{x}_t$ for $j \geq r+1$.

Proposition 2. *Let Assumptions 1–4 hold and assume $r+1 \leq \widehat{r} \leq \bar{r}$ for some fixed $\bar{r}$. Additionally, assume that $\mathbf{f}_t$ and $\boldsymbol{\varepsilon}_t$ are weakly stationary. Suppose*

$$\max_{1\leq i\leq n} \max_{r+1\leq j\leq \widehat{r}} |\widehat{w}_{x,ij}^{(\ell)}| = O_p(n^{-\alpha/2}), \tag{13}$$

for some $1 \leq \ell \leq L_T$ and $\alpha \in [0,1]$, and let $b_T = \log^{1/\beta+\delta} T$ for β in Assumption 4 and some fixed $\delta > 0$. Then,

$$\max_{1\leq i\leq n} \max_{1\leq t\leq T} |\widehat{\chi}_{it}^{\text{bpc}} - \chi_{it}| = O_p\left[n^{(1-\alpha)/2} \left(\sqrt{\frac{\log(n)}{T}} \vee \frac{1}{\sqrt{n}} \right) \log(T) \right]. \tag{14}$$

The proof of Proposition 2 is provided in Appendix A.2. Condition (13) is very general and its motivation is as follows. Writing $x_{it} = \sum_{j=1}^{n\wedge T} \widehat{w}_{x,ij} \widehat{\mathbf{w}}_{x,j}^\top \mathbf{x}_t$, we have

$$\frac{1}{T}\sum_{t=1}^{T} x_{it}^2 = \sum_{j=1}^{n\wedge T} \widehat{w}_{x,ij}^2 \widehat{\mu}_{x,j} < \infty \quad \text{a.s. for all } 1 \leq i \leq n, \tag{15}$$

which implies that $\max_{1\leq i\leq n} |\widehat{w}_{x,ij}| = O(\widehat{\mu}_{x,j}^{-1/2})$. In addition, the rate of convergence of the sample covariance matrix $n^{-1}\|\widehat{\boldsymbol{\Gamma}}_x - \boldsymbol{\Gamma}_x\| = O_p(\sqrt{\log(n)/T})$ (see Lemma 3) and (C4) yields $\widehat{\mu}_{x,j} = O_p(n\sqrt{\log(n)/T}) = o_p(n)$ for $j \geq r+1$. These arguments hold for blockwise estimators as well, and indicate that there may be (spuriously) large coordinates in the empirical eigenvectors $\widehat{\mathbf{w}}_{x,j}^{(\ell)}$, $j \geq r+1$ that fall in the regime of $\alpha < 1$. In other words, (13) is merely a consequence of the boundedness of $\mu_{x,j}$, $j \geq r+1$ without any further structural assumptions on the model (1).

It has been shown that for a random matrix $\mathbf{M} \in \mathbb{R}^{n\times T}$ with independent entries, the eigenvectors of $T^{-1}\mathbf{M}^\top\mathbf{M}$ are 'delocalised' in probability with the bound $1/\sqrt{n}$ up to a logarithmic factor (see Theorem B.3 of Vu and Wang (2015) and a survey given in O'Rourke, Vu and Wang (2016)). In view of this, when $\mathbf{x}_t \sim_{\text{iid}} (\mathbf{0}, \boldsymbol{\Gamma}_x)$ and follows an *exact* factor model with $\boldsymbol{\Gamma}_\varepsilon = \mathbf{I}_n$, the condition (13) is met with $\alpha = 1$ up to a logarithmic factor, and the consistency of the PC estimator derived in Theorem 1 carries over even with $\widehat{r} > r$. However, under the approximate time series factor model adopted in this paper, there is no such theoretical guarantee to the best of our knowledge. In Section 4, we verify that, under a variety of data generating models, the non-leading empirical eigenvectors indeed exhibit 'sparsity' with few very large coordinates, thus corresponding to the regime $\alpha \simeq 0$.

The following Examples 1–2 provide the lower bounds complementing upper bounds in (13) and (14) for a particular example where $\boldsymbol{\Gamma}_\varepsilon$ follows a sparse spiked covariance model. Together, Proposition 2 and Examples 1–2 are indicative of the potential pitfalls stemming from the over-estimation of r for the PC estimator of the common component.

Example 1 (Lower bound on $\max_{1\le i\le n}\max_{r+1\le j\le \widehat{r}}|\widehat{w}_{x,ij}|$)**.** We assume that $\mathbf{\Gamma}_\varepsilon = \Delta_n \mathbf{v}\mathbf{v}^\top + \sigma^2\mathbf{I}_n$ with $\|\mathbf{v}\|_0 \asymp n^\alpha$ for some $\alpha \in [0,1)$ and $\mathbf{v}^\top\mathbf{v} = 1$. Also, we suppose $n \asymp T^\kappa$ for some $\kappa > 0$ (see Assumption 2). This leads to

$$\mathbf{\Gamma}_x = \mathbf{W}_\chi \mathbf{M}_\chi \mathbf{W}_\chi^\top + \Delta_n \mathbf{v}\mathbf{v}^\top + \sigma^2\mathbf{I}_n, \tag{16}$$

where $\mathbf{W}_\chi$ is the $n \times r$ matrix of normalised eigenvectors and $\mathbf{M}_\chi$ the $r \times r$ diagonal matrix of eigenvalues of $\mathbf{\Gamma}_\chi$. Further, we assume that $\Delta_n \asymp n^\nu$ for some $\max(0, 1 - 1/(2\kappa) + \alpha/2) < \nu < 1$, and let $\mathbf{v}^\top\mathbf{w}_{\chi,j} = 0$ for all $j = 1,\dots,r$. In this model, the idiosyncratic component has a one-factor structure with a weakly pervasive factor where its 'strength' Δ_n increases with α. We may interpret this as the weak factor being prevalent in all the elements belonging to a group defined by the support of $\mathbf{v}$. In time series setting, such a structure has also been considered by De Mol, Giannone and Reichlin (2008), Lam, Yao and Bathia (2011) and Onatski (2012), among others.

When $\nu > 0$, the model (16) does not fulfil Assumption 1 (iv). However, even when $\nu \in (0,1)$, the oracle PC estimator obtained with $\widehat{r} = r$ can be shown to be consistent by adapting the proof of Proposition 1: From $\|\widehat{\mathbf{\Gamma}}_x - \mathbf{\Gamma}_\chi\| \le \|\widehat{\mathbf{\Gamma}}_x - \mathbf{\Gamma}_x\| + \|\mathbf{\Gamma}_\varepsilon\|$, we yield

$$\begin{aligned}
\|\widehat{\mathbf{W}}_x - \mathbf{W}_\chi\mathbf{S}\| &= O_p\left(\sqrt{\frac{\log(n)}{T}} \vee \frac{1}{n^{1-\nu}}\right), \quad \text{and} \\
\max_{1\le i\le n}\max_{1\le t\le T}|\widehat{\chi}_{it}^{\mathrm{pc}} - \chi_{it}| &= O_p\left\{\left(\sqrt{\frac{\log(n)}{T}} \vee \frac{1}{n^{(1-\nu)/2}}\right)\log(T)\right\}.
\end{aligned} \tag{17}$$

Under model (16), for large enough n, we have

$$\mathbf{w}_{x,j} = \begin{cases} \mathbf{w}_{\chi,j} & \text{for } 1 \le j \le r, \\ \mathbf{v} & \text{for } j = r+1, \end{cases} \quad \text{with } \mu_{x,j} = \begin{cases} \mu_{\chi,j} + \sigma^2 & \text{for } 1 \le j \le r, \\ \Delta_n + \sigma^2 & \text{for } j = r+1, \\ \sigma^2 & \text{for } r+2 \le j \le n. \end{cases}$$

As in (4), we apply Corollary 1 of Yu, Wang and Samworth (2015) and yield

$$\begin{aligned}
\|\widehat{\mathbf{w}}_{x,r+1} - s\mathbf{v}\| &\le \frac{2^{3/2}\|\widehat{\mathbf{\Gamma}}_x - \mathbf{\Gamma}_x\|}{\min(\mu_{x,r} - \mu_{x,r+1}, \mu_{x,r+1} - \mu_{x,r+2})} \\
&= O_p\left(n^{1-\nu}\sqrt{\frac{\log(n)}{T}}\right) = o_p(n^{-\alpha/2})
\end{aligned} \tag{18}$$

for some $s \in \{-1, 1\}$, i.e., $\widehat{\mathbf{w}}_{x,r+1}$ achieves consistency in estimating $\mathbf{v}$ albeit at a slower convergence rate than that reported in (4). Also, the sparsity of $\mathbf{v}$ leads to $n^{-\alpha/2}(\max_{1\le i\le n}|v_i|)^{-1} = O(1)$, and thus from (18), for some fixed $C_0 > 0$,

$$\max_{1\le i\le n}|\widehat{w}_{x,i,r+1}| = \max_{1\le i\le n}|v_i| + O_p\left(n^{1-\nu}\sqrt{\frac{\log(n)}{T}}\right) \ge C_0 n^{-\alpha/2}. \tag{19}$$

Example 2 (Lower bound on the estimation error in (14))**.** Continuing with the model (16) imposed on $\mathbf{\Gamma}_x$, we further assume that $\mathbf{x}_t \sim_{\text{iid}} \mathcal{N}_n(\mathbf{0}, \mathbf{\Gamma}_x)$ and $n \asymp T$ for simplicity (i.e., $\kappa = 1$) such that $\nu \in ((1+\alpha)/2, 1)$. Under independence, we simplify the blockwise estimator as

$$\widehat{\chi}_{it}^{\text{bpc}} = \sum_{j=1}^{\widehat{r}} \widehat{w}_{x,ij}^{(\ell)} (\widehat{\mathbf{w}}_{x,j}^{(\ell)})^\top \mathbf{x}_t \quad \text{for} \quad t \in I_\ell, \quad \ell = 0, 1,$$

with $I_0 = \{2u,\, 1 \le u \le \lfloor T/2 \rfloor\} = \bar{I}_1$ and $I_1 = \{2u+1,\, 0 \le u \le \lfloor T/2 \rfloor\} = \bar{I}_0$. Suppose that $\widehat{r} = r + 1$. Then there exist fixed $C_k > 0$, $1 \le k \le 4$ such that

$$\begin{aligned}
\max_{1\le i\le n} \max_{1\le t\le T} |\widehat{\chi}_{it}^{\text{bpc}} - \chi_{it}| &\ge \max_{\ell=0,1} \max_{t\in I_\ell} \left|\widehat{w}_{x,1,r+1}^{(\ell)}\right| \left|(\widehat{\mathbf{w}}_{x,r+1}^{(\ell)})^\top \mathbf{x}_t\right| \\
&\quad - \max_{1\le i\le n} \max_{\ell=0,1} \max_{t\in I_\ell} \left|\sum_{j=1}^{r} \widehat{w}_{x,ij}^{(\ell)} (\widehat{\mathbf{w}}_{x,j}^{(\ell)})^\top \mathbf{x}_t - \chi_{it}\right| \\
&\ge C_1 n^{-\alpha/2} \max_{\ell=0,1} \max_{t\in I_\ell} \left|(\widehat{\mathbf{w}}_{x,r+1}^{(\ell)})^\top \mathbf{x}_t\right| - C_2 n^{-(1-\nu)/2} \sqrt{\log(T)} \\
&\ge C_3 n^{-\alpha/2} \cdot \sqrt{\Delta_n \log(T)} - C_2 n^{\nu/2-1/2} \sqrt{\log(T)} \ge C_4 n^{(\nu-\alpha)/2} \sqrt{\log(T)}
\end{aligned}$$

where all the inequalities except for the first are understood as holding with probability tending to one. The second inequality follows from (17), (19) and that $n \asymp T$, with the rate $\sqrt{\log(T)}$ due to the stronger Gaussian assumption we impose here in place of the sub-exponential tail in Assumption 3 (ii). The penultimate inequality holds by Theorem 3.4 of Hartigan (2014) since for each $\ell = 0, 1$, we have $(\widehat{\mathbf{w}}_{x,r+1}^{(\ell)})^\top \mathbf{x}_t \sim_{\text{iid}} \mathcal{N}(0, \widetilde{\sigma}^2)$ for $t \in I_\ell$ with

$$\widetilde{\sigma}^2 \ge \Delta_n \left\{(\widehat{\mathbf{w}}_{x,r+1}^{(\ell)})^\top \mathbf{v}\right\}^2 + \sigma^2 \ge \Delta_n \left\{1 + o_p(n^{-\alpha/2})\right\} + \sigma^2$$

by applying Corollary 1 of Yu, Wang and Samworth (2015) as in (18). For comparison, we derive the upper bound on the estimation error in this setting as in Proposition 2. From (19), we have $\max_{\ell=0,1} \max_{1\le i\le n} |\widehat{w}_{x,i,r+1}^{(\ell)}| \asymp n^{-\alpha/2}$ and by adopting the arguments analogous to those used in the proof of Proposition 2, it is readily seen that

$$\max_{1\le i\le n} \max_{1\le t\le T} |\widehat{\chi}_{it}^{\text{bpc}} - \chi_{it}| = O_p(n^{(\nu-\alpha)/2} \sqrt{\log(T)}),$$

i.e., the lower bound matches the upper bound in terms of the rate.

Examples 1–2 demonstrate that in the presence of weak factors, the PC estimator can incur non-negligible error increasing with n due to the 'localised' behaviour of $\widehat{\mathbf{w}}_{x,j}$, $j \ge r+1$ when the factor number is over-estimated. In practice, such situations can emerge when the idiosyncratic component exhibits a group structure that induces the presence of weak factors. Chudik, Pesaran and Tosetti (2011) discuss the plausibility of *semi-weak* and *semi-strong* factors

corresponding to $\|\mathbf{\Gamma}_\varepsilon\| \asymp n^\nu$ with $\nu \in (0,1)$ in real datasets. In Section 5, the daily returns of the stocks comprising the Standard & Poor's 100 index are analysed where Figure 6 shows a clear group structure in the idiosyncratic component which is in line with the model (16). We also refer to Barigozzi and Hallin (2017) where the network structure in the idiosyncratic component of the similar dataset has been analysed in detail.

Remark 2 (Large covariance matrix estimation)**.** Fan, Lv and Qi (2011) and Fan, Liao and Mincheva (2013) investigate the problem of large covariance matrix estimation with an estimator comprised of a factor-driven covariance matrix of the common component and a thresholded idiosyncratic covariance matrix under the assumption of sparsity on $\mathbf{\Gamma}_\varepsilon$ (see (30)); in the former, the factors are assumed to be observable and the latter extends the estimator to the case of unobservable factors. For the consistency of the thresholded idiosyncratic covariance matrix, Assumption 2.2 of Fan, Lv and Qi (2011) requires $\widehat{\varepsilon}_{it}$, an estimator of ε_{it}, to satisfy

$$\max_{1\le i\le n} \frac{1}{T}\sum_{t=1}^{T} |\widehat{\varepsilon}_{it} - \varepsilon_{it}|^2 = o_p(1) \quad \text{and} \quad \max_{1\le i\le n}\max_{1\le t\le T} |\widehat{\varepsilon}_{it} - \varepsilon_{it}| = O_p(1), \tag{20}$$

and Lemma C.11 of Fan, Liao and Mincheva (2013) verifies the conditions for the PC estimator combined with the estimator of r proposed in Bai and Ng (2002). However, Proposition 1 indicates that the second condition in (20) may be violated when $\widehat{r} > r$. Moreover, our numerical studies in Section 4 demonstrate that neither of the conditions in (20) are fulfilled by the PC estimator when the idiosyncratic components are moderately correlated, which in turn implies that the covariance matrix estimator of Fan, Liao and Mincheva (2013) will suffer from relying on the accurate estimation of r. We further explore this point by applying our methodology to estimating the covariance of a panel of financial time series in Section 5.

3. Modification of the PC estimator

3.1. Scaled PC estimator

Recall that due to the presence of an eigengap (C3)–(C4) and the consistency of the r leading eigenvectors of $\widehat{\mathbf{\Gamma}}_x$ (see (4)), we obtain the uniform bound of $O_p(n^{-1/2})$ on $|\widehat{w}_{x,ij}|$, $j \le r$, see (7). In other words, with large probability, there exists some fixed $c_w > 0$ such that $|\widehat{w}_{x,ij}|$, $j \le r$ is bounded by $c_w n^{-1/2}$ uniformly in $1 \le i \le n$ and $1 \le j \le r$. Motivated by these observations, we propose the scaled PC estimator

$$\widehat{\chi}_{it}^{\mathrm{sc}} = \sum_{j=1}^{\widehat{r}} \widehat{w}_{x,ij}^{\mathrm{sc}} (\widehat{\mathbf{w}}_{x,j}^{\mathrm{sc}})^\top \mathbf{x}_t, \quad \text{where} \tag{21}$$

$$\widehat{\mathbf{w}}_{x,j}^{\mathrm{sc}} = \nu_j^{-1}\, \widehat{\mathbf{w}}_{x,j} \quad \text{with} \quad \nu_j = \max\left\{1, \frac{\sqrt{n}}{c_w} \max_{1\le i\le n} |\widehat{w}_{x,ij}|\right\}. \tag{22}$$

We can choose c_w such that with large probability, the proposed scaling does not alter the contribution from $\widehat{\mathbf{w}}_{x,j}$, $j \leq r$ to $\widehat{\chi}_{it}^{\text{sc}}$ by yielding $\nu_j = 1$ for $j \leq r$, even though it is applied *without* knowing r. On the contrary, for $\widehat{\mathbf{w}}_{x,j}$, $j \geq r+1$, any large contribution from the spurious factors is scaled down by the factor of ν_j.

Remark 3 (Choice of c_w)**.** In our numerical analysis, we observe that the performance of the scaled PC estimator does not vary much with respect to reasonably chosen c_w, see Figure 2 (the details of the experiment is deferred to Section C.5 of Barigozzi and Cho (2020)). Unlike e.g., the methods based on singular value thresholding, our scaled PC estimator does not 'kill' any factors including the spurious ones, and thus avoids the hazard of under-estimating the contribution of the factors completely provided that $\widehat{r} \geq r$. We find the choice of $c_w = 1.1 \times \sqrt{n}\ \max_{1 \leq i \leq n} |\widehat{w}_{x,i1}|$ works reasonably well and adopt it throughout the numerical studies, which is shown to work well for a range of models in Section 4.

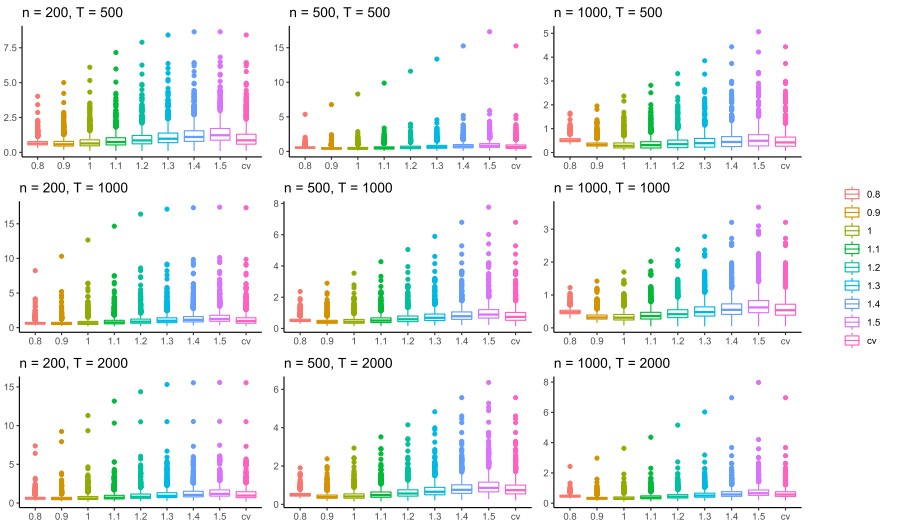

FIG 2. *Box plots of the estimation errors of* $\widehat{\chi}_{it}^{\text{bsc}}(c_w)$ *averaged over* 1000 *realisations generated under Model 1 of Section 4.1 with* $\phi = 1$, $n \in \{200, 500, 1000\}$ *(left to right) and* $T \in \{500, 1000, 2000\}$ *(top to bottom), for* $c_w \in \{0.8, \ldots, 1.5\} \times \sqrt{n}\ \max_{1 \leq i \leq n} |\widehat{w}_{x,i1}|$ *and the cross-validated choice ('CV') (left to right within each plot).*

Scaling preserves the orthogonality among $\widehat{\mathbf{w}}_{x,j}^{\text{sc}}$, $j \leq \widehat{r}$, which facilitates the theoretical treatment of the scaled PC estimator. Following the same reasoning as in Section 2.4, we continue the discussion on the theoretical properties of the scaled PC estimator by considering its blockwise counterpart, which, recalling

the notations from Section 2.4, is given by

$$\widehat{\chi}_{it}^{\text{bsc}} = \sum_{j=1}^{\widehat{r}} \widehat{w}_{x,ij}^{\text{sc},(\ell)} (\widehat{\mathbf{w}}_{x,j}^{\text{sc},(\ell)})^{\top} \mathbf{x}_t \quad \text{for} \quad t \in I_\ell, \quad 1 \le \ell \le L_T, \tag{23}$$

where $\widehat{\mathbf{w}}_{x,j}^{\text{sc},(\ell)}$ is defined analogously as in (22) with $\widehat{\mathbf{w}}_{x,j}^{(\ell)}$ in place of $\widehat{\mathbf{w}}_{x,j}$.

Proposition 3. *Let Assumptions 1–4 hold and suppose $r+1 \le \widehat{r} \le \bar{r}$ for some fixed $\bar{r}$. Additionally, assume that $\mathbf{f}_t$ and $\boldsymbol{\varepsilon}_t$ are weakly stationary. Then, there exists a fixed constant c_w satisfying*

$$c_w \ge \sqrt{n} \times \max_{1 \le \ell \le L_T} \max_{1 \le i \le n} \max_{1 \le j \le r} |\widehat{w}_{x,ij}|$$

such that

$$\max_{1 \le i \le n} \max_{1 \le t \le T} |\widehat{\chi}_{it}^{\text{bsc}} - \chi_{it}| = O_p \left\{ \left(\sqrt{\frac{\log(n)}{T}} \vee \frac{1}{\sqrt{n}} \right) \log(T) \right\}. \tag{24}$$

The proof is provided in Appendix A.3. Compared to Propositions 1 and 2, Proposition 3 establishes that under the same conditions, the scaled PC estimator attains the same rate of convergence as the oracle PC estimator obtained with the true number of factors, *without* requiring such knowledge and *regardless* of the behaviour of $\widehat{\mathbf{w}}_{x,j}$, $j \ge r+1$.

3.2. Relationship to capped PC estimator

Similarly motivated by the uniform boundedness of $|\widehat{w}_{x,ij}|$ for $j \le r$ (see (7)), Barigozzi, Cho and Fryzlewicz (2018) proposed the capped PC estimator of χ_{it}:

$$\widehat{\chi}_{it}^{\text{cp}} = \sum_{j=1}^{\widehat{r}} \widehat{w}_{x,ij}^{\text{cp}} (\widehat{\mathbf{w}}_{x,j}^{\text{cp}})^{\top} \mathbf{x}_t, \tag{25}$$

where each element of $\widehat{\mathbf{w}}_{x,j}^{\text{cp}}$ is obtained by capping $\widehat{w}_{x,ij}$ as

$$\widehat{w}_{x,ij}^{\text{cp}} = \widehat{w}_{x,ij} \, \mathbb{I}\left(|\widehat{w}_{x,ij}| \le \frac{c_w}{\sqrt{n}} \right) + \text{sign}(\widehat{w}_{x,ij}) \cdot \frac{c_w}{\sqrt{n}} \, \mathbb{I}\left(|\widehat{w}_{x,ij}| > \frac{c_w}{\sqrt{n}} \right) \tag{26}$$

for some fixed $c_w > 0$. Capping can be viewed as the projection of each $\widehat{\mathbf{w}}_{x,j}$ onto the ℓ_∞-sphere of radius $c_w n^{-1/2}$. As with scaling, asymptotically, capping does not alter the contribution from the leading r eigenvectors of $\widehat{\boldsymbol{\Gamma}}_x$, while it truncates any large contribution from spurious factors when $\widehat{r} \ge r+1$, all *without* the knowledge of the true r. We generalise Theorem 2 of Barigozzi, Cho and Fryzlewicz (2018) whereby lifting the assumption of Gaussianity imposed on $\boldsymbol{\varepsilon}_t$ in the latter; the proof can be found in Section B.3 of Barigozzi and Cho (2020).

Proposition 4. *Let Assumptions 1–4 hold and suppose $r+1 \le \widehat{r} \le \bar{r}$ for some fixed $\bar{r}$. Then, there exists a fixed constant c_w satisfying $c_w \ge \sqrt{n} \times \max_{1 \le i \le n} \max_{1 \le j \le r} |\widehat{w}_{x,ij}|$ such that $\max_{1 \le i \le n} \max_{1 \le t \le T} |\widehat{\chi}_{it}^{\text{cp}} - \chi_{it}| = O_p(\log T)$.*

Refinement of the upper bound given in Proposition 4 is a difficult task as reasoned in (a)–(b) of Section 2.4, even when considering its blockwise version, $\widehat{\chi}_{it}^{\mathrm{bcp}}$, due to the lack of orthogonality of the capped eigenvectors. Nevertheless, Proposition 4 shows that the capped estimator $\widehat{\chi}_{it}^{\mathrm{cp}}$ improves upon the worst case performance of the PC estimator reported in (11).

Unlike the capped PC estimator, the scaled PC estimator shrinks down the eigenvectors after modification from $\|\widehat{\mathbf{w}}_{x,j}\|^2 = 1$ to $\|\widehat{\mathbf{w}}_{x,j}^{\mathrm{sc}}\|^2 = \nu_j^{-1}$, which further curtails the spurious contribution from $\widehat{\mathbf{w}}_{x,j}$, $j \geq r+1$ as demonstrated in the following example.

Example 3. For simplicity, let us ignore the stochastic nature of $\widehat{\mathbf{w}}_{x,j}$ and suppose that $\widehat{\mathbf{w}}_{x,j'}$ for some $j' \geq r+1$ is approximately sparse. That is, there exists $\mathcal{S} \subset \{1,\ldots,n\}$ with $|\mathcal{S}| = O(1)$ and a fixed $c_0 > 0$ such that $|\widehat{w}_{x,ij'}| \geq c_0$, $i \in \mathcal{S}$, while $\max_{i \notin \mathcal{S}} |\widehat{w}_{x,ij'}| = O(n^{-1/2})$. Then, we have $\|\widehat{\mathbf{w}}_{x,j'}^{\mathrm{sc}}\|^2 \leq c_w(c_0\sqrt{n})^{-1}$, which shrinks the overall contribution of the j'-th estimated factor to $\widehat{\boldsymbol{\chi}}_t^{\mathrm{sc}}$ by the factor of $\sqrt{n}$, in comparison with that to the PC estimator. In the same scenario, however, capping does not always lead to $\|\widehat{\mathbf{w}}_{x,j'}^{\mathrm{cp}}\| = o(1)$. Consider e.g., $\widehat{\mathbf{w}}_{x,j'} = (1/\sqrt{2}, 1/\sqrt{2(n-1)}, \ldots, 1/\sqrt{2(n-1)})^\top$ and $c_w/\sqrt{n} \geq 1/\sqrt{2(n-1)}$, in which case $\|\widehat{\mathbf{w}}_{x,j'}^{\mathrm{cp}}\| \geq 1/\sqrt{2}$.

3.3. Relationship to eigenvalue shrinkage

Recalling that $\max_{1\leq i\leq n} |\widehat{w}_{x,ij}| = O(\widehat{\mu}_{x,j}^{-1/2})$ (see the discussion following (15)), we may re-write the scaling factor ν_j using the choice

$$c_w = 1.1 \times \sqrt{n} \max_{1\leq i\leq n} |\widehat{w}_{x,i1}|$$

as suggested in Remark 3:

$$\nu_j = \max\left\{1, \frac{\max_{1\leq i\leq n}|\widehat{w}_{x,ij}|}{1.1 \times \max_{1\leq i\leq n}|\widehat{w}_{x,i1}|}\right\} = \max\left\{1, \sqrt{\frac{C_j\widehat{\mu}_{x,1}}{\widehat{\mu}_{x,j}}}\right\}, \quad \text{such that}$$

$$\widehat{\chi}_{it}^{\mathrm{sc}} = \sum_{j=1}^{\widehat{r}} \min\left\{1, \sqrt{\frac{\widehat{\mu}_{x,j}}{C_j\widehat{\mu}_{x,1}}}\right\} \widehat{w}_{x,ij}\widehat{\mathbf{w}}_{x,j}^\top \mathbf{x}_t$$

with some fixed $C_j > 0$. In other words, for some choice of c_w, the scaled PC estimator admits a representation as a PC estimator combined with the eigenvalue-based shrinkage. Ideal choices for C_j are $C_j \gg \widehat{\mu}_{x,j}/\widehat{\mu}_{x,1}$ for $j \geq r+1$, and $C_j \leq \widehat{\mu}_{x,j}/\widehat{\mu}_{x,1}$ for $j \leq r$ which, however, are infeasible since they require the knowledge of r. We consider a simpler but feasible choice of $C_j = 1$ for all j, and define the modified PC estimator based on eigenvalue shrinkage:

$$\widehat{\chi}_{it}^{\mathrm{sh}} = \sum_{j=1}^{\widehat{r}} \sqrt{\frac{\widehat{\mu}_{x,j}}{\widehat{\mu}_{x,1}}} \cdot \widehat{w}_{x,ij}(\widehat{\mathbf{w}}_{x,j})^\top \mathbf{x}_t. \tag{27}$$

Its blockwise version $\widehat{\chi}_{it}^{\mathrm{bsh}}$ is defined analogously with $\widehat{\mathbf{w}}_{x,j}^{(\ell)}$ and the corresponding eigenvalues $\widehat{\mu}_{x,j}^{(\ell)}$ replacing $\widehat{\mathbf{w}}_{x,j}$ and $\widehat{\mu}_{x,j}$, respectively. This estimator is ex-

pected to keep under control the over-estimation error, since $\widehat{\mu}_{x,j}/\widehat{\mu}_{x,1} = o_p(1)$ for $j \geq r+1$ while being asymptotically bounded away from zero for $j \leq r$. Hence, $\widehat{\chi}_{it}^{\text{sh}}$ preserves the contribution of the leading PCs although with a possible bias. From the simulation studies in Section 4, we observe that any bias incurred by over-shrinkage is well compensated by its effectiveness in shrinking down the spuriously large over-estimation error.

Remark 4 (Eigenvalue shrinkage)**.** The good performance of the shrinkage estimator in (27) may be explained by its link to the literature on eigenvalue shrinkage-based estimators. Donoho, Gavish and Johnstone (2018) and Donoho and Ghorbani (2018) investigate the optimal eigenvalue shrinkage for spiked covariance matrix estimation when $\mathbf{x}_t \sim_{\text{iid}} \mathcal{N}_n(\mathbf{0}, \mathbf{\Gamma}_x)$ with $\mu_{x,1} \geq \ldots \geq \mu_{x,r} > 1$ and $\mu_{x,j} = 1$, $j \geq r+1$. It has been shown that for any loss function considered therein, the optimal eigenvalue shrinkage function η yields $\eta(\widehat{\mu}_{x,j}) < \widehat{\mu}_{x,j}$. Heuristically, shrinkage of eigenvalues not only accounts for the upward shift of empirical eigenvalues, but also the inconsistency in empirical eigenvectors (Donoho, Gavish and Johnstone, 2018).

4. Simulation studies

4.1. Set-up

We consider the following data generating model which allows for serial correlations in f_{jt} and both serial and cross-sectional correlations in ε_{it}.

$$x_{it} = r^{-1/2} \sum_{j=1}^{r} \lambda_{ij} f_{jt} + \sqrt{\phi}\varepsilon_{it},\ 1 \leq i \leq n;\ 1 \leq t \leq T, \quad \text{where} \tag{28}$$
$$f_{jt} = \rho_{f,j} f_{j,t-1} + u_{jt}, \quad \varepsilon_{it} = \rho_{\varepsilon,i}\varepsilon_{i,t-1} + v_{it},$$

with factor loadings $\lambda_{ij} \sim_{\text{iid}} \mathcal{N}(0,1)$, factor innovations $u_{jt} \sim_{\text{iid}} \mathcal{N}(0, 1/(1-\rho_{f,j}^2))$, and the autoregressive parameters as $\rho_{f,j} = \rho_f - 0.05(j-1)$ with $\rho_f = 0.5$. For the idiosyncratic innovations v_{it}, we consider the following two models.

Model 1. With $H = 10$, $e_{it} \sim_{\text{iid}} \mathcal{N}(0, 1-\rho_{\varepsilon,i}^2)$ and $\beta_i \sim_{\text{iid}} \text{Unif}\{-0.15, 0.15\}$, we generate $v_{it} = (1 + 2\beta_i^2 H)^{-1/2}(e_{it} + \beta_i \sum_{l=i-H, l\neq i}^{i+H} e_{lt})$, and set $\rho_{\varepsilon,i} \sim_{\text{iid}} \text{Unif}\{0.2, -0.2\}$. This model has been taken from Bai and Ng (2002) except that we select the parameters $\rho_{\varepsilon,i}$, β_i and H of smaller magnitude such that the problem of identifying r is in fact *easier* here than in the original paper.

Model 2 (Cai, Ma and Wu, 2015). The vector $\mathbf{v}_t = (v_{1t}, \ldots, v_{nt})^\top$ is such that $\mathbf{v}_t = \mathbf{\Gamma}_v^{1/2}\mathbf{e}_t$, where $\mathbf{\Gamma}_v = \mathbf{V}\mathbf{\Delta}\mathbf{V}^\top + \mathbf{I}_n$, and $\mathbf{e}_t \sim_{\text{iid}} \mathcal{N}_n(0, (1-\rho_{\varepsilon,i}^2)\mathbf{I}_n)$. The diagonal matrix $\mathbf{\Delta}$ has r non-zero eigenvalues taking equidistant values from 20 to 10, and $\mathbf{V}$ is chosen as the r leading left singular vectors of a matrix $\mathbf{M} \in \mathbb{R}^{n\times r}$, whose first $[\varrho n]$ rows are drawn independently from $\mathcal{N}(0,1)$ and the rest are set to zero. By construction, this models adds r additional 'weak' factors stemming from the large (although bounded for all n) eigenvalues of $\mathbf{\Gamma}_v$.

Model 1 provides a benchmark as it is popularly adopted in the factor model literature, while Model 2 mimics the case of weak factors considered in Examples 1–2.

We control the 'noise-to-signal' ratio with $\phi \in \{0.5, 1, 2\}$ such that larger values of ϕ correspond to the low signal-to-noise ratio. Throughout, we set $r = 5$, and consider $T \in \{500, 1000, 2000\}$ and $n \in \{200, 500, 1000\}$, and $\varrho \in \{0.2, 0.5, 0.9\}$ for Model 2.

We explore the in-sample estimation accuracy of the PC estimator $\widehat{\chi}_{it}^{\mathrm{pc}}$ in (2), the scaled estimator $\widehat{\chi}_{it}^{\mathrm{sc}}$ in (21), the capped estimator $\widehat{\chi}_{it}^{\mathrm{cp}}$ in (25) and the shrinkage estimator $\widehat{\chi}_{it}^{\mathrm{sh}}$ in (27), with and without blockwise estimation for which we set $b_T = [\log^2 T]$. For estimating r, we consider the two estimators (8) ('BN') and (9) ('AH'), setting $r_{\max} = [\sqrt{n \wedge T}]$. For comparison, we also investigate the performance of the oracle estimator $\widehat{\chi}_{it}^{\mathrm{oracle}}$ defined as the PC estimator (2) obtained with the true r. For each setting, 1000 realisations have been generated. We provide the results obtained under Model 1 in the main text, and additional simulation results under Model 2 are available in Section C of Barigozzi and Cho (2020). Based on Figure 1, we report the results when the factor number estimator (8) is used, in order to contrast the behaviour of the proposed modified PC estimators to that of the PC estimator in terms of their 'insensitivity' to the over-estimation of r.

4.2. Results

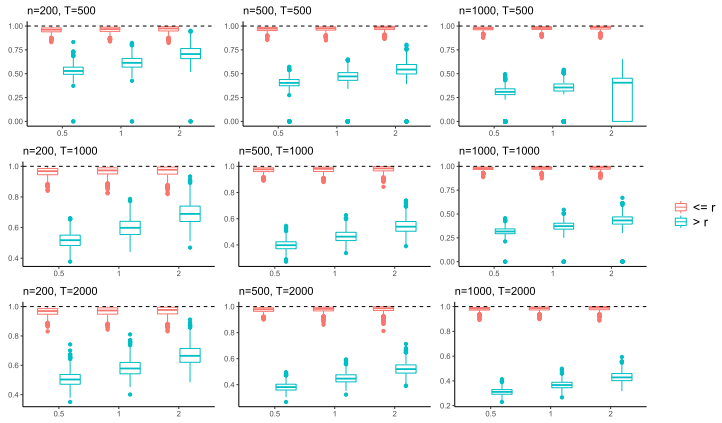

FIG 3. *Box plots of* $\|\widehat{\mathbf{w}}_{x,j}^{\mathrm{sc}}\|$ *averaged for* $2 \le j \le r$ *('*$\le r$*') against that averaged for* $r+1 \le j \le \widehat{r}$ *('*$> r$*') averaged* 1000 *realisations generated under Model 1 with* $n \in \{200, 500, 1000\}$ *(left to right),* $T \in \{500, 1000, 2000\}$ *(top to bottom) and* $\phi \in \{0.5, 1, 2\}$ *(left to right within each plot).*

First, we investigate the amount of scaling and capping applied to $\widehat{\mathbf{w}}_{x,j}$ when $j \le r$ and $j \ge r+1$, in order to verify whether the asymptotic argument in (7) is valid for finite n and T. Figure 3 plots the norm of the scaled eigenvectors $\|\widehat{\mathbf{w}}_{x,j}^{\mathrm{sc}}\|$

in (22) averaged over $2 \leq j \leq r$ and $r+1 \leq j \leq \widehat{r}$, respectively, for varying T, n and ϱ. The results confirm that across different scenarios, scaling does not alter the contribution from the leading r eigenvectors of $\widehat{\mathbf{\Gamma}}_x$, while curtailing that from $\widehat{\mathbf{w}}_{x,j}$, $j \geq r+1$ by yielding $\|\widehat{\mathbf{w}}^{\mathrm{sc}}_{x,j}\| \ll \|\widehat{\mathbf{w}}_{x,j}\| = 1$, $j \geq r+1$ especially for large n. This in turn indicates that there are a few spuriously large coordinates in $\widehat{\mathbf{w}}_{x,j}$, $j \geq r+1$, corresponding to the regime $\alpha \simeq 0$ in Proposition 2. As shown below, this leads to the undesirable behaviour of the PC estimator while affecting the modified PC estimators to a much lesser degree.

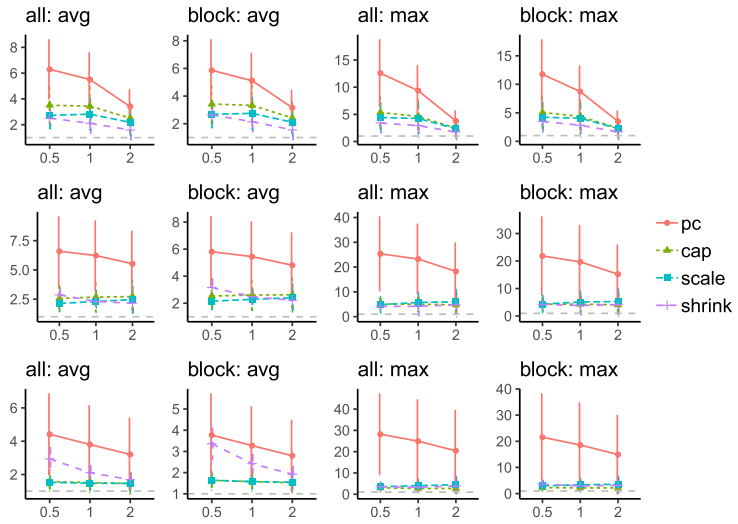

FIG 4. *$\mathrm{err}_{\mathrm{avg}}(\widehat{\chi}^{\circ}_{it})$ and $\mathrm{err}_{\max}(\widehat{\chi}^{\circ}_{it})$ of $\widehat{\chi}^{\mathrm{pc}}_{it}$, $\widehat{\chi}^{\mathrm{cp}}_{it}$, $\widehat{\chi}^{\mathrm{sc}}_{it}$ and $\widehat{\chi}^{\mathrm{sh}}_{it}$ estimated using the entire sample ('all'), and their blockwise counterparts ('block'), averaged over* 1000 *realisations generated under Model 1 with $T = 500$, $n \in \{200, 500, 1000\}$ (top to bottom) and $\phi \in \{0.5, 1, 2\}$ (left to right within each plot). The vertical errors bars represent the standard deviations.*

Next, we evaluate the accuracy of an estimator $\widehat{\chi}^{\circ}_{it}$ of χ_{it} relative to that of the oracle estimator, using the following error measures

$$\mathrm{err}_{\mathrm{avg}} = \frac{n^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}(\widehat{\chi}^{\circ}_{it} - \chi_{it})^2}{\widehat{\mathsf{E}}\{n^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}(\widehat{\chi}^{\mathrm{oracle}}_{it} - \chi_{it})^2\}},$$

$$\mathrm{err}_{\max} = \frac{\max_{1\leq i\leq n}\sum_{t=1}^{T}(\widehat{\chi}^{\circ}_{it} - \chi_{it})^2}{\widehat{\mathsf{E}}\{\max_{1\leq i\leq n}\sum_{t=1}^{T}(\widehat{\chi}^{\mathrm{oracle}}_{it} - \chi_{it})^2\}},$$

where $\widehat{\mathsf{E}}$ denotes the average over all Monte Carlo repetitions, and $\circ$ denotes the use of PC, capped, scaled or shrinkage estimator and their blockwise counterparts. We note that $\mathrm{err}_{\max}$ is specifically to capture the possible deterioration in the estimators for individual i due to the over-estimation of r, and both measures are closely related to the conditions in (20). Figures 4–5 show the average and standard deviation of $\mathrm{err}_{\mathrm{avg}}$ and $\mathrm{err}_{\max}$ over 1000 Monte Carlo realisations.

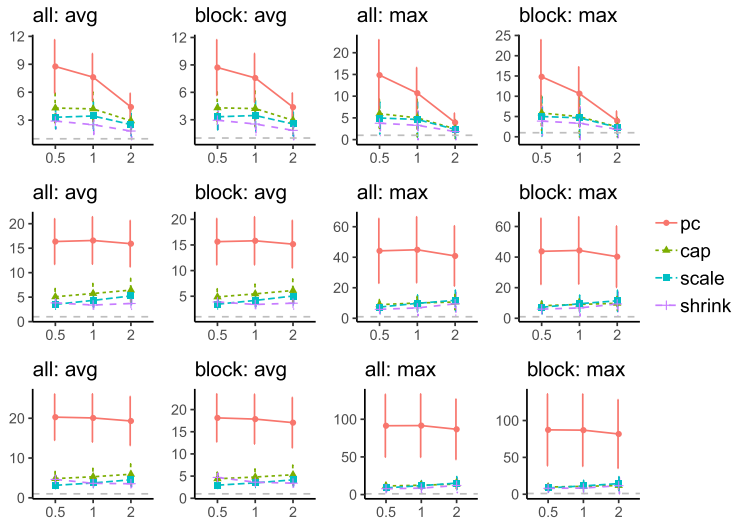

FIG 5. *$err_{\text{avg}}(\widehat{\chi}_{it}^{\circ})$ and $err_{\max}(\widehat{\chi}_{it}^{\circ})$ under Model 1 with $T = 2000$.*

Overall, blockwise estimators do not lose efficiency compared to their whole sample counterparts or, even perform slightly better in terms of the relative efficiency compared to the oracle PC estimator. It is evident that PC estimator exhibits the worst performance in almost all cases, in terms of both the average and variability of the two different error measures. Indeed, $\text{err}_{\max}^{\text{pc}}$ indicates that the PC estimator with an over-estimated factor number can be worse by hundredfold than the oracle PC estimator for some coordinates.

Capping and scaling lead to considerable improvement with respect to both error measures, with and without blockwise estimation, and marginally the scaled PC estimator tends to return smaller estimation errors. We note that $\widehat{\chi}_{it}^{\text{sh}}$ yields the smallest estimation error in many scenarios. Exceptions occur when n is relatively larger than T: the PC-based estimator of the factor space is expected to be highly accurate in this setting due to the blessing of dimensionality, and the bias introduced by eigenvalue shrinkage tends to deteriorate the performance of $\widehat{\chi}_{it}^{\text{sh}}$ (see Figure C.8 in Barigozzi and Cho (2020)).

As the signal-to-noise ratio decreases, the gap between the performance of $\widehat{\chi}_{it}^{\text{pc}}$ and our modified estimators gets closer, as the consistent estimation of χ_{it} itself becomes more challenging, i.e., the error due to the over-estimation of r in (10) becomes dominated by the first term. Increasing n also tends to close this gap as the performance of the estimator of r improves. This, however, has the opposite effect on $\text{err}_{\max}(\widehat{\chi}_{it}^{\text{pc}})$ since the maximum is taken over the n cross-sections. In general, err_{avg} and $\text{err}_{\max}$ evaluated at modified PC estimators exhibit much less fluctuations as n and T vary. Noting the close relationship between $\text{err}_{\text{avg}}(\widehat{\chi}_{it}^{\text{pc}})$ and $\text{err}_{\max}(\widehat{\chi}_{it}^{\text{pc}})$ and the conditions in (20), the results reported here indicate that the popularly adopted covariance matrix estimator based on factor analysis will suffer from the over-estimation of r.

5. Real data analysis

We consider risk minimisation for a portfolio consisting of the log returns of the daily closing values of the stocks comprising the Standard & Poor's 100 (S&P100) index between July 2006 and September 2013 (denoted by $\{x_{it},\, i \leq n,\, t \leq T\}$ with $n = 90$ and $T = 1814$) following the exercise conducted in Lam (2016). The dataset is available from Yahoo Finance.

As evidence of structural changes has been observed in a similar financial data dataset (Barigozzi, Cho and Fryzlewicz, 2018), we choose to adopt a rolling window of size $\widetilde{T} = 253$ (number of trading days each year) and evaluate the performance of a portfolio on a monthly basis (with 21 as the number of trading days each month). At the beginning of each month, different methods are adopted to estimate the covariance matrix of stock returns using one year of past returns. Each portfolio has weights given by

$$\widehat{\boldsymbol{\omega}}_k^{\circ} = \underset{\boldsymbol{\omega}\in\mathbb{R}^n:\, \boldsymbol{\omega}^\top \mathbf{1}_n = 1}{\arg\min}\, \boldsymbol{\omega}^\top \widehat{\boldsymbol{\Sigma}}_k^{\circ} \boldsymbol{\omega} = \frac{\widehat{\boldsymbol{\Sigma}}_k^{\circ} \mathbf{1}_n}{\mathbf{1}_n^\top \widehat{\boldsymbol{\Sigma}}_k^{\circ} \mathbf{1}_n} \quad \text{for } k = 1, \ldots, M = \lceil (T - \widetilde{T})/21 \rceil,$$

where $\widehat{\boldsymbol{\Sigma}}_k^{\circ}$ denotes some covariance matrix estimator based on the k-th rolling window $R_k = [21(k-1)+1, 21(k-1)+\widetilde{T}]$, and $\mathbf{1}_n$ denotes a vector consisting of n ones. At the end of each month, we compute the total excess return, the out-of-sample variance and the mean Sharpe ratio, given by

$$\widehat{\tau}(\widehat{\boldsymbol{\Sigma}}^{\circ}) = \sum_{k=1}^{M} \sum_{t=21(k-1)+1}^{\min(21k, T-\widetilde{T})} (\widehat{\boldsymbol{\omega}}_k^{\circ})^\top \mathbf{x}_{\widetilde{T}+t},$$

$$\widehat{\sigma}^2(\widehat{\boldsymbol{\Sigma}}^{\circ}) = \frac{1}{T-\widetilde{T}} \sum_{k=1}^{M} \sum_{t=21(k-1)+1}^{\min(21k, T-\widetilde{T})} \{(\widehat{\boldsymbol{\omega}}_k^{\circ})^\top \mathbf{x}_{\widetilde{T}+t} - \widehat{\mu}(\widehat{\boldsymbol{\Sigma}}_k^{\circ})\}^2 \quad \text{and}$$

$$\mathrm{SR}(\widehat{\boldsymbol{\Sigma}}^{\circ}) = \frac{1}{M} \sum_{k=1}^{M} \frac{\widehat{\tau}(\widehat{\boldsymbol{\Sigma}}_k^{\circ})}{\widehat{\sigma}(\widehat{\boldsymbol{\Sigma}}_k^{\circ})},$$

where $\widehat{\tau}(\widehat{\boldsymbol{\Sigma}}_k^{\circ})$, $\widehat{\mu}(\widehat{\boldsymbol{\Sigma}}_k^{\circ})$ and $\widehat{\sigma}^2(\widehat{\boldsymbol{\Sigma}}_k^{\circ})$ denote the total and mean excess return and the out-of-sample variance calculated for each portfolio from $\widehat{\boldsymbol{\Sigma}}_k^{\circ}$.

For covariance matrix estimation, we consider the following two approaches that separately estimate the factor-driven and idiosyncratic contributions.

Exact factor modelling (EFM). We force the covariance matrix of the idiosyncratic component to be diagonal and obtain

$$\widehat{\boldsymbol{\Sigma}}_k^{\mathrm{pc}} = \widetilde{T}^{-1} \sum_{t\in R_k} (\widehat{\boldsymbol{\chi}}_t^{\mathrm{pc}})^\top \widehat{\boldsymbol{\chi}}_t^{\mathrm{pc}} + \mathsf{diag}\Big(\widetilde{T}^{-1} \sum_{t\in R_k} (\widehat{\boldsymbol{\varepsilon}}_t^{\mathrm{pc}})^\top \widehat{\boldsymbol{\varepsilon}}_t^{\mathrm{pc}}\Big), \tag{29}$$

where the operator $\mathsf{diag}(\mathbf{A})$ returns a diagonal matrix with the diagonal elements of $\mathbf{A}$ in its diagonal, $\widehat{\boldsymbol{\chi}}_t^{\mathrm{pc}}$, $t \in R_k$ is obtained from $\widehat{\boldsymbol{\Gamma}}_{x,k} =$

$\widetilde{T}^{-1}\sum_{t\in R_k}\mathbf{x}_t\mathbf{x}_t^\top$ and $\widehat{\boldsymbol{\varepsilon}}_t^{\mathrm{pc}} = \mathbf{x}_t - \widehat{\boldsymbol{\chi}}_t^{\mathrm{pc}}$. Similarly, we yield $\widehat{\boldsymbol{\Sigma}}_k^{\mathrm{cp}}$, $\widehat{\boldsymbol{\Sigma}}_k^{\mathrm{sc}}$ and $\widehat{\boldsymbol{\Sigma}}_k^{\mathrm{sh}}$ with $\widehat{\boldsymbol{\chi}}_t^{\mathrm{cp}}$, $\widehat{\boldsymbol{\chi}}_t^{\mathrm{sc}}$ and $\widehat{\boldsymbol{\chi}}_t^{\mathrm{sh}}$ replacing $\widehat{\boldsymbol{\chi}}_t^{\mathrm{pc}}$ in (29), respectively. Also, we consider their blockwise versions $\widehat{\boldsymbol{\Sigma}}_k^{\mathrm{bpc}}$, $\widehat{\boldsymbol{\Sigma}}_k^{\mathrm{bcp}}$, $\widehat{\boldsymbol{\Sigma}}_k^{\mathrm{bsc}}$ and $\widehat{\boldsymbol{\Sigma}}_k^{\mathrm{bsh}}$ with the size of blocks $b = \lceil \log^2 \widetilde{T} \rceil$.

POET (Fan, Liao and Mincheva, 2013). We adopt the POET estimator

$$\widehat{\boldsymbol{\Sigma}}_k^{\mathrm{poet}} = \widetilde{T}^{-1}\sum_{t\in R_k}(\widehat{\boldsymbol{\chi}}_t^{\mathrm{pc}})^\top\widehat{\boldsymbol{\chi}}_t^{\mathrm{pc}} + \mathcal{T}\Big(\widetilde{T}^{-1}\sum_{t\in R_k}(\widehat{\boldsymbol{\varepsilon}}_t^{\mathrm{pc}})^\top\widehat{\boldsymbol{\varepsilon}}_t^{\mathrm{pc}}\Big), \tag{30}$$

where $\mathcal{T}(\cdot)$ performs an element-wise hard-thresholding (except for its diagonals) with an adaptively chosen threshold recommended by Fan, Liao and Mincheva (2013) including a constant selected via cross-validation.

The results obtained for varying $\widehat{r}$ are reported in Table 1; note that $\widehat{r} = 6$ is selected by the information criterion of Bai and Ng (2002) applied to the whole data. We note that $\widehat{r} = 2$ may already be over-estimating the number of factors, in that the idiosyncratic component estimated with $\widehat{r} = 1$ exhibits a prominent group structure and this may be falsely detected as a factor via PCA, see Figure 6. As demonstrated in Example 1, possibly highly structured nature of the idiosyncratic component may lead to some elements of $\widehat{\mathbf{w}}_{x,j}$, $j \geq 2$ being (spuriously) large, and poor performance of the corresponding PC estimator (see also Proposition 2 and Example 2). The EFM-based method combined with the PC estimator with $\widehat{r} = 2$ performs the best among all the methods. With $\widehat{r} = 6$, the POET yields large negative total excess returns and large out-of-sample variance, which confirms our observation in Remark 2 that the covariance (precision) matrix estimation based on factor modelling is susceptible to the errors arising from factor number estimation. Overall, the methods based on EFM perform better than the POET according to all measures considered. Among the modified PC estimators, the one that applies the largest amount of shrinkage ($\widehat{\boldsymbol{\chi}}_t^{\mathrm{sh}}$) achieves the most consistent performance with regards to the choice of $\widehat{r}$. Interestingly, the blockwise approach yields marginally better performance than the corresponding whole sample counterparts.

TABLE 1
Performance of portfolios constructed with different covariance estimators.

method		$\widehat{r}=2$			$\widehat{r}=4$			$\widehat{r}=6$		
		$\widehat{\tau}$	$\widehat{\sigma}^2$	SR	$\widehat{\tau}$	$\widehat{\sigma}^2$	SR	$\widehat{\tau}$	$\widehat{\sigma}^2$	SR
EFM	$\widehat{\boldsymbol{\chi}}_t^{\mathrm{pc}}$	25.032	0.917	0.932	-28.321	0.936	0.003	-22.515	0.818	0.460
	$\widehat{\boldsymbol{\chi}}_t^{\mathrm{cp}}$	22.513	0.922	0.865	-31.045	0.883	-0.028	-21.691	0.753	0.439
	$\widehat{\boldsymbol{\chi}}_t^{\mathrm{sc}}$	20.164	0.900	0.888	-24.468	0.832	0.133	-27.461	0.742	0.301
	$\widehat{\boldsymbol{\chi}}_t^{\mathrm{sh}}$	14.072	0.890	0.809	4.079	0.861	0.652	4.618	0.835	0.703
	$\widehat{\boldsymbol{\chi}}_t^{\mathrm{bpc}}$	21.288	0.835	0.907	-17.115	0.813	0.331	-14.934	0.750	0.468
	$\widehat{\boldsymbol{\chi}}_t^{\mathrm{bcp}}$	20.807	0.842	0.888	-19.331	0.818	0.270	-14.538	0.756	0.478
	$\widehat{\boldsymbol{\chi}}_t^{\mathrm{bsc}}$	20.626	0.833	0.886	-16.133	0.804	0.367	-13.939	0.760	0.469
	$\widehat{\boldsymbol{\chi}}_t^{\mathrm{bsh}}$	18.866	0.828	0.859	10.638	0.825	0.766	10.605	0.822	0.777
POET		-76.446	6.061	-0.315	-299.643	69.678	-0.331	-355.691	85.428	0.254

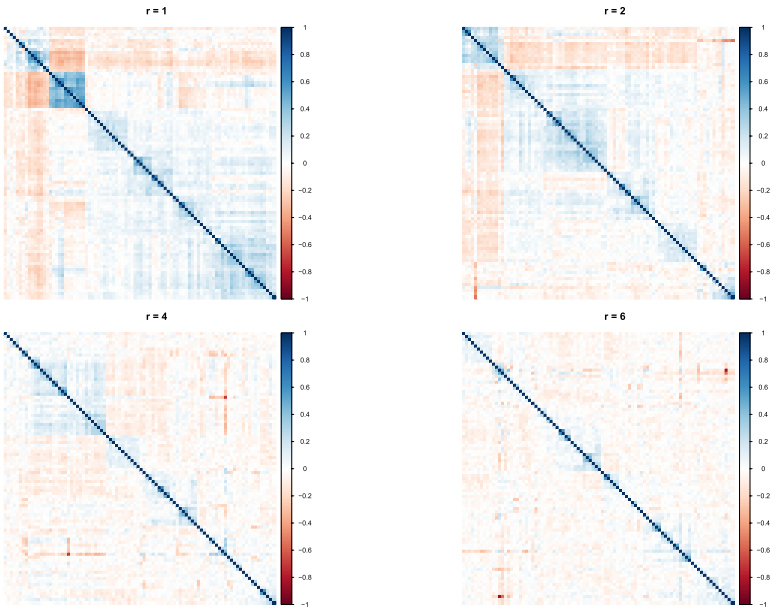

FIG 6. *Heatmap of the correlation matrix of the idiosyncratic component estimated with* $\widehat{r} \in \{1, 2, 4, 6\}$ *(from left to right, top to bottom). The variables are ordered via hierarchical clustering with the complete linkage method.*

6. Conclusions

Factor number estimation is a challenging task due to the lack of a clear gap in empirical eigenvalues, and various estimators tend to over-estimate r in the presence of moderate cross-sectional correlations in the idiosyncratic component. In this paper, we make the first attempt at establishing the non-negligibility of the error due to the over-estimation of r in the widely adopted PC estimator. In doing so, we propose a novel blockwise estimation technique, which enables rigorous treatment of this over-estimation error under a time series factor model. Also, we propose the modified PC estimators which are easily implemented and perform as well as the oracle PC estimator with obtained with r known, and verify this via extensive simulation studies. In practice, we recommend the use of $\widehat{\chi}_{it}^{\text{sh}}$ unless n is much greater than T (an unlikely setting for e.g., economic and financial data), which shows very good practical performance both on simulated and real-life datasets.

Appendix A: Proofs

A.1. Preliminaries

The following lemmas hold under Assumptions 1–4. Their proof can be found in Section B.1 of Barigozzi and Cho (2020).

Lemma 1. $\max_{1\le i\le n}\max_{1\le t\le T}|\chi_{it}| = O_p(\log(T))$, $\max_{1\le i\le n}\max_{1\le t\le T}|\varepsilon_{it}| = O_p(\log(T))$, and $\max_{1\le i\le n}\max_{1\le t\le T}|x_{it}| = O_p(\log(T))$.

Lemma 2. Let b_T satisfy $b_T\to\infty$ and $T^{-1}b_T\to 0$, and $L_T = \lceil T/b_T\rceil$.

(i) $\max_{1\le\ell\le L_T} n^{-1}\|\widehat{\mathbf{\Gamma}}_x^{(\ell)} - \mathbf{\Gamma}_\chi\| = O_p\Big(\sqrt{\frac{\log(n)}{T}}\vee\frac{1}{n}\Big)$.

(ii) $\max_{1\le\ell\le L_T}\max_{1\le i\le n} n^{-1/2}\|\boldsymbol{\varphi}_i^\top(\widehat{\mathbf{\Gamma}}_x^{(\ell)} - \mathbf{\Gamma}_\chi)\| = O_p\left(\sqrt{\frac{\log(n)}{T}}\vee\frac{1}{\sqrt{n}}\right)$.

Also, there exists an orthonormal matrix $\mathbf{S}_\ell\in\mathbb{R}^{r\times r}$ which, for $\widehat{\mathbf{W}}_x^{(\ell)} = [\widehat{\mathbf{w}}_{x,j}^{(\ell)},\ j\le r]$, satisfies

(iii) $\max_{1\le\ell\le L_T}\|\widehat{\mathbf{W}}_x^{(\ell)} - \mathbf{W}_\chi\mathbf{S}_\ell\| = O_p\left(\sqrt{\frac{\log(n)}{T}}\vee\frac{1}{n}\right)$;

(iv) $\max_{1\le\ell\le L_T}\max_{1\le i\le n}\sqrt{n}\,\|\boldsymbol{\varphi}_i^\top(\widehat{\mathbf{W}}_x^{(\ell)} - \mathbf{W}_\chi\mathbf{S}_\ell)\| = O_p\left(\sqrt{\frac{\log(n)}{T}}\vee\frac{1}{\sqrt{n}}\right)$.

Lemma 3.

(i) $n^{-1}\|\widehat{\mathbf{\Gamma}}_x - \mathbf{\Gamma}_\chi\| = O_p\Big(\sqrt{\frac{\log(n)}{T}}\vee\frac{1}{n}\Big)$.

(ii) $\max_{1\le i\le n} n^{-1/2}\|\boldsymbol{\varphi}_i^\top(\widehat{\mathbf{\Gamma}}_x - \mathbf{\Gamma}_\chi)\| = O_p\left(\sqrt{\frac{\log(n)}{T}}\vee\frac{1}{\sqrt{n}}\right)$.

(iii) $\max_{1\le i\le n}\sqrt{n}\,\|\boldsymbol{\varphi}_i^\top(\widehat{\mathbf{W}}_x - \mathbf{W}_\chi\mathbf{S})\| = O_p\left(\sqrt{\frac{\log(n)}{T}}\vee\frac{1}{\sqrt{n}}\right)$ for some orthonormal $r\times r$ matrix $\mathbf{S}$.

Lemma 4. *Let $\ell(t)$ denote the index of the block for which $t\in I_{\ell(t)}$, and $b_T = \log^{1/\beta+\delta}T$ for some $\delta>0$. Then,* $\max_{1\le t\le T}|(\widehat{\mathbf{w}}_{x,j}^{\ell(t)})^\top\boldsymbol{\varepsilon}_t| = O_p(\log(T))$.

A.2. Proof of Proposition 2

Recall the definition of $\ell(t)$ in Lemma 4. Note that

$$\max_{1\le i\le n}\max_{1\le t\le T}|\widehat{\chi}_{it}^{\text{bpc}} - \chi_{it}| \le \max_{1\le i\le n}\max_{1\le t\le T}\Big|\sum_{j=1}^{r}\widehat{w}_{x,ij}^{\ell(t)}(\widehat{\mathbf{w}}_{x,j}^{\ell(t)})^\top\mathbf{x}_t - \chi_{it}\Big|$$
$$+\max_{1\le i\le n}\max_{1\le t\le T}\Big|\sum_{j=r+1}^{\widehat{r}}\widehat{w}_{x,ij}^{\ell(t)}(\widehat{\mathbf{w}}_{x,j}^{\ell(t)})^\top\boldsymbol{\chi}_t\Big| + \max_{1\le i\le n}\max_{1\le t\le T}\Big|\sum_{j=r+1}^{\widehat{r}}\widehat{w}_{x,ij}^{\ell(t)}(\widehat{\mathbf{w}}_{x,j}^{\ell(t)})^\top\boldsymbol{\varepsilon}_t\Big|$$
$$=: I + II + III.$$

From Lemmas 1 and 2 (iii)–(iv), using the analogous arguments as those adopted in the proof of Proposition 1 in Section B.2 of Barigozzi and Cho (2020),

$$I \le \max_{1\le i\le n}\max_{1\le t\le T}|\boldsymbol{\varphi}_i^\top\widehat{\mathbf{W}}_x^{\ell(t)}(\widehat{\mathbf{W}}_x^{\ell(t)})^\top\mathbf{x}_t - \boldsymbol{\varphi}_i^\top\mathbf{W}_\chi\mathbf{S}_{\ell(t)}(\widehat{\mathbf{W}}_x^{\ell(t)})^\top\mathbf{x}_t|$$
$$+\max_{1\le i\le n}\max_{1\le t\le T}|\boldsymbol{\varphi}_i^\top\mathbf{W}_\chi\mathbf{S}_{\ell(t)}(\widehat{\mathbf{W}}_x^{\ell(t)})^\top\mathbf{x}_t - \boldsymbol{\varphi}_i^\top\mathbf{W}_\chi\mathbf{W}_\chi^\top\mathbf{x}_t|$$
$$+\max_{1\le i\le n}\max_{1\le t\le T}|\boldsymbol{\varphi}_i^\top\mathbf{W}_\chi\mathbf{W}_\chi^\top\boldsymbol{\varepsilon}_t|$$

$$=O_p\left\{\left(\sqrt{\frac{\log(n)}{T}}\vee\frac{1}{\sqrt{n}}\right)\log(T)\right\}. \tag{31}$$

Let $\widehat{\mathbf{W}}^{\ell(t)}_{x,(r+1):k}=[\widehat{\mathbf{w}}^{\ell(t)}_{x,j},\, r+1\le j\le k]$. Under Assumption 1 (i), it follows that $\mathbf{W}_\chi^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{W}_\chi=\mathbf{M}_\chi$ and hence $\mathbf{W}_\chi$ may be regarded as the left singular vectors of $\mathbf{\Lambda}$. Then, from the orthogonality of eigenvectors, (C1) and Lemma 2 (iii),

$$\begin{aligned}&\max_{1\le\ell\le L_T}\|(\widehat{\mathbf{W}}^{(\ell)}_{x,(r+1):k})^\top\mathbf{\Lambda}\|=\max_{1\le\ell\le L_T}\|(\widehat{\mathbf{W}}^{(\ell)}_{x,(r+1):k})^\top\mathbf{W}_\chi\mathbf{M}_\chi^{1/2}\|\\&\le\max_{1\le\ell\le L_T}\|(\widehat{\mathbf{W}}^{(\ell)}_{x,(r+1):k})^\top(\mathbf{W}_\chi\mathbf{S}_\ell-\widehat{\mathbf{W}}^{(\ell)}_{x})\|\,\|\mathbf{M}_\chi^{1/2}\|=O_p\left(\sqrt{\frac{n\log(n)}{T}}\vee\frac{1}{\sqrt{n}}\right)\end{aligned} \tag{32}$$

for any fixed $k\ge r+1$. Together with the condition (13) and Lemma 1, it yields

$$\begin{aligned}II&=O_p\left\{n^{-\alpha/2}\cdot\left(\sqrt{\frac{n\log(n)}{T}}\vee\frac{1}{\sqrt{n}}\right)\cdot\log(T)\right\}\\&=O_p\left\{\left(\sqrt{\frac{n^{1-\alpha}\log(n)}{T}}\vee\sqrt{\frac{1}{n^{(1+\alpha)}}}\right)\log(T)\right\}.\end{aligned}$$

Finally, from Lemma 4, $III=O_p(n^{-\alpha/2}\log(T))$, and the conclusion follows.

A.3. Proof of Proposition 3

Recall the definition of $\ell(t)$ in Lemma 4. Note that

$$\begin{aligned}\max_{1\le i\le n}\max_{1\le t\le T}|\widehat{\chi}_{it}^{\mathrm{bsc}}-\chi_{it}|&\le\max_{1\le i\le n}\max_{1\le t\le T}\Big|\sum_{j=1}^{r}\widehat{w}^{\mathrm{sc},\ell(t)}_{x,ij}(\widehat{\mathbf{w}}^{\mathrm{sc},\ell(t)}_{x,j})^\top\mathbf{x}_t-\chi_{it}\Big|\\&\quad+\max_{1\le i\le n}\max_{1\le t\le T}\Big|\sum_{j=r+1}^{\widehat{r}}\widehat{w}^{\mathrm{sc},\ell(t)}_{x,ij}(\widehat{\mathbf{w}}^{\mathrm{sc},\ell(t)}_{x,j})^\top\boldsymbol{\chi}_t\Big|\\&\quad+\max_{1\le i\le n}\max_{1\le t\le T}\Big|\sum_{j=r+1}^{\widehat{r}}\widehat{w}^{\mathrm{sc},\ell(t)}_{x,ij}(\widehat{\mathbf{w}}^{\mathrm{sc},\ell(t)}_{x,j})^\top\boldsymbol{\varepsilon}_t\Big|\\&=:I+II+III.\end{aligned}$$

Since scaling does not alter the r leading eigenvectors with probability tending to one, thanks to the arguments leading to (7) and Lemma 2, we derive that $I=O_p\{(\sqrt{\log(n)/T}\vee 1/\sqrt{n})\log(T)\}$ as in (31). Next, due to the orthogonality of $\widehat{\mathbf{w}}^{\mathrm{sc},\ell(t)}_{x,j}$, $j\le\widehat{r}$,

$$II=\max_{1\le i\le n}\max_{1\le t\le T}\Big|\sum_{j=r+1}^{\widehat{r}}\widehat{w}^{\mathrm{sc},\ell(t)}_{x,ij}(\widehat{\mathbf{w}}^{\mathrm{sc},\ell(t)}_{x,j})^\top\{\boldsymbol{\chi}_t-\widehat{\mathbf{W}}^{\ell(t)}_{x,1:r}(\widehat{\mathbf{W}}^{\ell(t)}_{x,1:r})^\top\mathbf{x}_t\}\Big|$$

$$= O_p\left\{\left(\sqrt{\frac{\log(n)}{T}} \vee \frac{1}{\sqrt{n}}\right)\log(T)\right\}$$

from the bound on I and the uniform boundedness of $|\widehat{w}_{x,ij}^{\mathrm{sc},\ell(t)}|$. Finally, Lemma 4 and the definition of $|\widehat{w}_{x,ij}^{\mathrm{sc},\ell(t)}|$ yield $III = O_p(\log(T)/\sqrt{n})$, which concludes the proof.

References

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81** 1203–1227. MR3064065

Alessi, L., Barigozzi, M. and Capasso, M. (2010). Improved penalization for determining the number of factors in approximate static factor models. *Statistics and Probability Letters* **80** 1806–1813. MR2734245

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171. MR1956857

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259

Bai, J. and Ng, S. (2007). Determining the number of primitive shocks in factor models. *Journal of Business & Economic Statistics* **25** 52–60. MR2338870

Bai, J. and Ng, S. (2019). Rank regularized estimation of approximate factor models. *Journal of Econometrics* **212** 78–96. MR3994008

Barigozzi, M., Cho, H. and Fryzlewicz, P. (2018). Simultaneous multiple change-point and factor analysis for high-dimensional time series. *Journal of Econometrics* **206** 187–225. MR3840788

Barigozzi, M. and Cho, H. (2020). Consistent estimation of high-dimensional factor models when the factor number is over-estimated. *arXiv preprint* arXiv:1811.00306.

Barigozzi, M. and Hallin, M. (2017). A network analysis of the volatility of high dimensional financial series. *Journal of the Royal Statistical Society: Series C* **66** 581–605. MR3632343

Cai, T. T., Ma, Z. and Wu, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics* **41** 3074–3110. MR3161458

Cai, T. T., Ma, Z. and Wu, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields* **161** 781–815. MR3334281

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51** 1281–1304. MR0736050

Chudik, A., Pesaran, M. H. and Tosetti, E. (2011). Weak and strong cross-section dependence and estimation of large panels. *Econometrics Journal* **14** 45–90. MR2797084

De Mol, C., Giannone, D. and Reichlin, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics* **146** 318–328. MR2465176

Donoho, D. L., Gavish, M. and Johnstone, I. M. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *The Annals of Statistics* **46** 1742–1778. MR3819116

Donoho, D. L. and Ghorbani, B. (2018). Optimal covariance estimation for condition number loss in the spiked model. *arXiv preprint* arXiv:1810.07403.

Doz, C., Giannone, D. and Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics* **164** 188–205. MR2821802

El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics* **36** 2757–2790. MR2485012

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B* **75** 603–680. MR3091653

Fan, J. and Liao, Y. (2019). Learning latent factors from diversified projections and its applications to over-estimated and weak factors. *arXiv preprint* arXiv:1908.01252.

Fan, J., Lv, J. and Qi, L. (2011). Sparse high dimensional models in economics. *Annual Review of Economics* **3** 291–317.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000). The Generalized Dynamic Factor Model: identification and estimation. *The Review of Economics and Statistics* **82** 540–554.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2005). The Generalized Dynamic Factor Model: one-sided estimation and forecasting. *Journal of the American Statistical Association* **100** 830–840. MR2201012

Forni, M., Giannone, D., Lippi, M. and Reichlin, L. (2009). Opening the black box: structural factor models versus structural VARs. *Econometric Theory* **25** 1319–1347. MR2540502

Hartigan, J. (2014). Bounding the maximum of dependent random variables. *Electronic Journal of Statistics* **8** 3126–3140. MR3301303

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics* **29** 295–327. MR1863961

Lam, C. (2016). Nonparameteric eigenvalue-regularized precision or covariance matrix estimation. *The Annals of Statistics* **44** 928–953. MR3485949

Lam, C., Yao, Q. and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika* **98** 901–918. MR2860332

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics* **40** 694–726. MR2933663

Merlevède, F., Peligrad, M. and Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields* **151** 435–474. MR2851689

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics* **92** 1004–1016.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*

168 244–258. MR2923766

Onatski, A. (2015). Asymptotic analysis of the squared estimation error in misspecified factor models. *Journal of Econometrics* **186** 388–406. MR3343793

O'Rourke, S., Vu, V. and Wang, K. (2016). Eigenvectors of random matrices: a survey. *Journal of Combinatorial Theory: Series A* **144** 361–442. MR3534074

Paul, D. and Aue, A. (2014). Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference* **150** 1–29. MR3206718

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* **97** 1167–1179. MR1951271

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics* **20** 147–162. MR1963257

Ting, C.-M., Ombao, H., Samdin, S. B. and Salleh, S.-H. (2017). Estimating dynamic connectivity states in fMRI using regime-switching factor models. *IEEE Trans Med Imaging* **37** 1011–1023.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B* **61** 611–622. MR1707864

Trapani, L. (2018). A randomized sequential procedure to determine the number of factors. *Journal of the American Statistical Association* **113** 1341–1349. MR3862361

Vu, V. and Wang, K. (2015). Random weighted projections, random quadratic forms and random eigenvectors. *Random Structures & Algorithms* **47** 792–821. MR3418916

Yu, L., He, Y. and Zhang, X. (2018). Robust factor number specification for large-dimensional factor model. *arXiv preprint* arXiv:1808.09107. MR4001652

Yu, Y., Wang, T. and Samworth, R. J. (2015). A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika* **102** 315–323. MR3371006